Numerical Weather Prediction

The Met Office GCR(k) Algorithm

Forecasting Research Technical Report No. 395

Timothy Payne

email: nwp_publications@metoffice.com

©Crown Copyright

The Met Office GCR(k) Algorithm

T. J. Payne

October 2002

Abstract

The Generalised Conjugate Residual (GCR) method, designated GCR(k) if restarted every k iterations, is a Krylov subspace projection method originally developed by Eisenstat, Elman and Schultz (EES) in 1983 to solve the equation

$$L\mathbf{x} = \mathbf{f}$$

for $\mathbf{x}$ where the matrix L is nonsymmetric but has definite symmetric part. A version of it is used in the New Dynamics and in the Perturbation Forecast (PF) and Adjoint Models in 4D-Var to solve a Helmholtz-type equation for the pressure correction. The version we use in the Met Office (hereafter MO GCR(k)) and by Smolarkiewicz et al differs from EES's in the way it is restarted, thereby altering some characteristics of the method.

In this report we analyse the properties of MO GCR(k) and apply the results to the PF and Adjoint Models.

Contents

1 Introduction 4

2 MO GCR(k) and Related Algorithms 5

3 Petrov-Galerkin-Krylov Algorithms 6
- 3.1 Projection Method (Petrov-Galerkin) . . . 7
- 3.2 Choice of basis vectors for $\mathcal{K}_m$ . . . 8
- 3.3 Restarts and truncations . . . 8
- 3.4 Restarted/Truncated Petrov-Galerkin-Krylov Algorithm RT-PGK(b, θ, ϕ) . . . 8
- 3.5 MO GCR, Standard (EES) GCR, GMRES and Orthomin as Special Cases of RT-PGK . . . 9

4 Link with Damped Wave Equation 9

5 Properties of The Iterates $\mathbf{r}_i$, $\mathbf{p}_i$ in MO GCR(k) 10
- 5.1 Relations between the Iterates in The MO GCR(k) . . . 11
- 5.2 Failure of MO GCR(k) . . . 12
- 5.3 Convergence of the Modulii of the Residuals of MO GCR(k) . . . 12
- 5.4 Convergence of the Difference of the Residuals of MO GCR(k) . . . 12
- 5.5 Homogeneity of MO GCR(k) . . . 12
- 5.6 Stationary State of MO GCR(k) . . . 13
- 5.7 The Asymptotic Behaviour of the MO GCR(k) . . . 13

6 Conditions for Convergence of The MO GCR(k) Algorithm 14
- 6.1 Minimisation Property and Convergence of MO GCR(k) . . . 14
- 6.2 Convergence of MO GCR(k) For Large k . . . 16

7 Geometric perspective: MO GCR(k) as an Iterated Map 16
- 7.1 Example: MO GCR(1) . . . 17

8 Rate of Convergence of MO GCR(k) and Choice of Preconditioner 19
- 8.1 Relations between the iterates if LP^{-1} is symmetric . . . 19
- 8.2 Rate of Convergence of MO GCR(k) for Symmetric LP^{-1} . . . 21
- 8.3 Preconditioning GCR(k) to Accelerate Convergence . . . 23

9 Cost of the MO GCR(k) 23

10 The Eigenvalues of L in the PF and Adjoint Models 24

11 The Eigenvalues of LP^{-1} and the rate of convergence of MO GCR(k) in the PF and Adjoint Models 28

12 Summary 31

Appendices **32**

A Discretising The Damped Wave Equation (15) with k=3 yields Orthomin(2) 32

B Proofs of Results in Section 5 34

C Calculating the Eigenvalues of L and LP^{-1} in the PF and Adjoint Models 35
- C.1 Extremal Eigenvalues of L . . . 35
- C.2 Extremal Eigenvalues of LP^{-1} . . . 37
- C.3 Smallest Eigenvalues of L . . . 37
- C.4 Smallest Eigenvalues of LP^{-1} . . . 38

D The Preconditioned MO GCR(k) Algorithm 39

E Note on the code and settings used for the test runs 40

References **40**

Acknowledgement: Useful discussions were held with several members of the New Dynamics Group at the Met Office, in particular Terry Davies, Mark Mawson and Nigel Wood.

1 Introduction

A version of the Generalised Conjugate Residual (GCR) Method is used to solve the Helmholtz-type equation for the pressure correction in the New Dynamics and in the PF and Adjoint Models in 4D-Var.

GCR, designated GCR(k) if restarted every k iterations, is a Krylov subspace projection method developed by Eisenstat, Elman and Schultz (EES) in 1983 to solve the equation

$$L\mathbf{x} = \mathbf{f} \tag{1}$$

for $\mathbf{x}$ where the matrix L is nonsymmetric but has definite symmetric part. The version we use in the Met Office (hereafter MO GCR(k)) and by Smolarkiewicz et al differs from EES's in the way it is restarted, thereby altering some characteristics of the method.

In Section 3 we show that MO GCR(k) can be placed within the unifying framework of Petrov-Galerkin-Krylov algorithms proposed by Saad and Schultz in 1985. This enables comparison between it and related methods such as EES's GCR(k), Orthomin(k) and GMRES(k). Smolarkiewicz and Margolin showed in [10] how Orthomin(1) (which coincides with MO GCR(1)) may be derived as a steady-state solution to the discretisation of a damped second order wave equation. It has been suggested in the literature that a similar discretisation of a (k+1)th order damped wave equation would yield MO GCR(k). In Section 4 we show that in fact it yields Orthomin(k), but that it is possible to obtain MO GCR(k) by discretising a cycle of damped wave equations of orders 2,..,k+1.

In Sections 5 and 6 we derive the main properties of the algorithm. For Met Office applications GCR(k) is used with a preconditioner P (block vertical or ADI), ie effectively we solve $LP^{-1}P\mathbf{x}=\mathbf{f}$ for $P\mathbf{x}$. As with EES's GCR(k) MO GCR(k) will converge for any positive integer k as long as the symmetric part of LP^{-1} is definite. Other convergence criteria are presented. In Section 7 we formulate MO GCR(k) as an iterated map on $R^{(k+1)n}$ which affords a geometric view of its properties including convergence and non-convergence.

Analysis of the convergence rate of MO GCR(k) is facilitated by the fact that in our applications LP^{-1} is close to symmetric. In Section 8 it is shown that in the exactly symmetric case the iterates generated by MO GCR(k) are independent of k. This is used to estimate MO GCR(k)'s convergence rate, which as for the classical conjugate gradient algorithm is largely determined by the distribution of eigenvalues. Preconditioning improves this eigenvalue distribution.

The material of Sections 2 to 8 concerns MO GCR(k) as applied to any linear equation of the form (1). Thereafter most of the material relates to applying this variety of GCR(k) to the Helmholtz-type equation of the PF and Adjoint Models in 4D-Var.

In Section 10 we use Arnoldi-type methods (described in Appendix C) to calculate the eigenvalues of L in the PF and Adjoint Models. The ellipticity parameter G which appears in the denominator of some coefficients of L is of the form

$$1 + (\Delta t)^2 \frac{\partial \theta}{\partial z}$$

so evidently will go negative for large enough Δt at any point of static instability. At a timestep where at some point on the grid G passes through zero an eigenvalue of L passes through infinity, and in particular L is indefinite for all sufficiently large timesteps. If a positive lower limit is imposed on G then L is definite for all timesteps.

In Section 11 the eigenvalues of LP^{-1} are calculated and the results of Section 8 used with some success to predict from them the convergence rate of preconditioned GCR(k).

2 MO GCR(k) and Related Algorithms

Orthomin(k) and the variants of GCR(k) are truncations or restarts of the basic GCR algorithm [2] to solve (1) for $\mathbf{x}$[1]:

Given an initial guess $\mathbf{x}_0$, set $\mathbf{r}_0$=$\mathbf{f} - L\mathbf{x}_0$ ānd $\mathbf{p}_0 = \mathbf{r}_0$
For i=0 until converged do:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + a_i \mathbf{p}_i \tag{2}$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i - a_i L\mathbf{p}_i \tag{3}$$

where

$$a_i = \frac{\langle \mathbf{r}_i, L\mathbf{p}_i \rangle}{\langle L\mathbf{p}_i, L\mathbf{p}_i \rangle} \tag{4}$$

$$\mathbf{p}_{i+1} = \mathbf{r}_{i+1} + b_0^i \mathbf{p}_0 + ... + b_i^i \mathbf{p}_i \tag{5}$$

where

$$b_j^i = -\frac{\langle L\mathbf{r}_{i+1}, L\mathbf{p}_j \rangle}{\langle L\mathbf{p}_j, L\mathbf{p}_j \rangle} \tag{6}$$

End do

(4) means that a_i is chosen so that $|\mathbf{r}_{i+1}|$ is minimised subject to being of the form (3), (6) means that $\{b_j^i$,j=0,..,i$\}$ are chosen so that for all i

$$\langle L\mathbf{p}_i, L\mathbf{p}_j \rangle = 0 \text{ any } j \neq i \tag{7}$$

Plainly the algorithm fails if for some i $L\mathbf{p}_i = \mathbf{0}$, see Section 5.2. We shall assume where necessary that $L\mathbf{p}_i \neq \mathbf{0}$. Note (by virtue of the initial choice of $\mathbf{r}_0$, (2,3) and induction on i) that for all i

$$\mathbf{r}_i = \mathbf{f} - L\mathbf{x}_i \tag{8}$$

Orthomin(k) [14], the 'original' GCR(k) of [2], and the version of GCR(k) used at the Met Office (identical to the version in [11]), referred to here as MO GCR(k), are restarted or truncated versions of GCR as follows:

MO GCR(k) is:

Given an initial guess $\mathbf{x}_0$, set $\mathbf{r}_0$=$\mathbf{f} - L\mathbf{x}_0$ and $\mathbf{p}_0 = \mathbf{r}_0$

For j=1 until converged do:
For i=1 to k do:
Calculate in order a_{i-1}, $\mathbf{x}_i$, $\mathbf{r}_i$, $(b_0^{i-1}, b_1^{i-1}, ..., b_{i-1}^{i-1})$, $\mathbf{p}_i$
End do
Set $\mathbf{r}_0 \leftarrow \mathbf{r}_k$, $\mathbf{p}_0 \leftarrow \mathbf{p}_k$, $\mathbf{x}_0 \leftarrow \mathbf{x}_k$
End do

A different way of restarting at the end of each loop, which we will refer to as **EES GCR(k)** because it was first described in [2], is as follows:

[1] Below and for most of this report we follow the convention of [2] in setting $\mathbf{r}_i = \mathbf{f} - L\mathbf{x}_i$. The Met Office code uses $\mathbf{r}_i = L\mathbf{x}_i - \mathbf{f}$ which leads to some changes of sign, see listing in Appendix D.

Given an initial guess $\mathbf{x}_0$, set $\mathbf{r}_0$=$\mathbf{f} - L\mathbf{x}_0$ and $\mathbf{p}_0 = \mathbf{r}_0$

For j=1 until converged do:
Calculate in order a_0, $\mathbf{x}_1$, $\mathbf{r}_1$
For i=1 to k do:
Calculate in order $(b_0^{i-1}, b_1^{i-1},..., b_{i-1}^{i-1})$, $\mathbf{p}_i$, a_i, $\mathbf{x}_{i+1}$, $\mathbf{r}_{i+1}$
End do
Set $\mathbf{r}_0 \leftarrow \mathbf{r}_{k+1}$, $\mathbf{p}_0 \leftarrow \mathbf{r}_{k+1}$, $\mathbf{x}_0 \leftarrow \mathbf{x}_{k+1}$
End do

A third variety of GCR is **Orthomin(k)**, which is the same as MO GCR(k) for the first k iterations:

Given an initial guess $\mathbf{x}_0$, set $\mathbf{r}_0$=$\mathbf{f} - L\mathbf{x}_0$ and $\mathbf{p}_0 = \mathbf{r}_0$

For i=1 until converged do
Calculate in order a_{i-1}, $\mathbf{x}_i$, $\mathbf{r}_i$,
$(b_{max[i-k,0]}^{i-1}, b_{max[i-k,0]+1}^{i-1},..., b_{i-1}^{i-1})$, $\mathbf{p}_i$
End do

where for i≥k (5) is replaced by

$$\mathbf{p}_{i+1} = \mathbf{r}_{i+1} + b_{i-k}^i \mathbf{p}_{i-k} + ... + b_i^i \mathbf{p}_i \quad [Orthomin(k+1)]$$

Note:

(1) With MO GCR(k) at the end of each loop $\mathbf{p}_0 \leftarrow \mathbf{p}_k$ unlike in EES GCR(k) where $\mathbf{p}_0 \leftarrow \mathbf{r}_{k+1}$. Thus the EES GCR(k) begins each loop with a complete restart while the MO GCR(k) carries over $\mathbf{p}$ from the previous loop

(2) The MO GCR(1) is equivalent to Orthomin(1) and is known generally as the conjugate residual method (eg [10]). After k-1 "set-up" steps each Orthomin(k) iteration has the form

Calculate in order a_{i-1},$\mathbf{x}_i$,$\mathbf{r}_i$,$(b_{i-k}^{i-1}, b_{i-k+1}^{i-1},...,b_{i-1}^{i-1})$,$\mathbf{p}_i$

MO GCR(k) is equivalent to cycles of such iterations, as follows:

Given an initial guess $\mathbf{x}_0$, set $\mathbf{r}_0$=$\mathbf{f} - L\mathbf{x}_0$ and $\mathbf{p}_0 = \mathbf{r}_0$

For j=1 until converged do:
Orthomin(1) iteration, Orthomin(2) iteration,...,Orthomin(k) iteration
Set $\mathbf{r}_0 \leftarrow \mathbf{r}_k$, $\mathbf{p}_0 \leftarrow \mathbf{p}_k$, $\mathbf{x}_0 \leftarrow \mathbf{x}_k$
End do

(3) It is easily seen that the relation (8) is still true for MO GCR(k). Thus $\mathbf{x}_i$ converges to the required solution if and only if the residuals (or errors) $\mathbf{r}_i \rightarrow \mathbf{0}$.

3 Petrov-Galerkin-Krylov Algorithms

In [8] Saad and Schultz present a unifying framework for the conjugate-gradient like algorithms used for nonsymmetric linear systems. We present a version of this scheme and show how MO GCR(k), EES GCR(k), Orthomin(k) and GMRES(k) are special cases of it.

Suppose the equation to solve is (1) where L is a (probably large and sparse) nonsymmetric n × n matrix.

3.1 Projection Method (Petrov-Galerkin)

Let $\mathcal{K}$, $\mathcal{M}$ be m-dimensional subspaces of R^n. If $\mathbf{x}_0$ is an initial guess for (1) then $\tilde{\mathbf{x}}$ satisfies the Petrov-Galerkin conditions if

$$\tilde{\mathbf{x}} \in \mathbf{x}_0 + \mathcal{K} \tag{9}$$

$$\mathbf{f} - L\tilde{\mathbf{x}} \perp \mathcal{M} \tag{10}$$

Because of the following two propositions, proved in [7], there are two sensible choices for $\mathcal{M}$, $\mathcal{M} = \mathcal{K}$ and $\mathcal{M}$=L$\mathcal{K}$.

Proposition ([7], Proposition 5.2) If $\mathcal{M} = \mathcal{K}$ and L is symmetric positive definite then $\tilde{\mathbf{x}}$ satisfies (9,10) iff it minimises

$$\langle L(\mathbf{x} - \mathbf{x}_*), \mathbf{x} - \mathbf{x}_* \rangle$$

over

$$\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}$$

where $\mathbf{x}_*$ is the solution to (1).

Proposition ([7], Proposition 5.3) If $\mathcal{M} = L\mathcal{K}$ then $\tilde{\mathbf{x}}$ satisfies (9,10) iff it minimises

$$|\mathbf{f} - L\mathbf{x}|$$

over

$$\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}$$

It is usually more convenient to work with

$$\mathbf{r} = \mathbf{f} - L\mathbf{x},\ \mathbf{r}_0 = \mathbf{f} - L\mathbf{x}_0$$

than with $\mathbf{x}$, $\mathbf{x}_0$, so (9,10) become

$$\tilde{\mathbf{r}} \in \mathbf{r}_0 + L\mathcal{K},\ \tilde{\mathbf{r}} \perp \mathcal{M} \tag{11}$$

The above propositions suggest the following as a generic Petrov-Galerkin (PG) method for solving (1), in which

$$\mathcal{K} = span\{\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{m-1}\}$$

with m increasing by 1 each iteration.

1. *Initialise* : *Guess* $\mathbf{x}_0$, *Set* $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{f} - L\mathbf{x}_0$
   Do steps 2 and 3 for m=0,1,2,..
2. *Set* $\mathbf{x}_{m+1} = \mathbf{x}_m - \sum_{i=0}^{m} \alpha_i \mathbf{p}_i,\ \mathbf{r}_{m+1} = \mathbf{r}_m + \sum_{i=0}^{m} \alpha_i L\mathbf{p}_i$
   where α_i are chosen by imposing the PG conditions
   $\mathbf{r}_{m+1} \perp \mathbf{p}_j,\ j = 0, 1, .., m$ (*corresponding to choosing* $\mathcal{M} = \mathcal{K}$)
   or
   $\mathbf{r}_{m+1} \perp L\mathbf{p}_j,\ j = 0, 1, .., m$ (*corresponding to choosing* $\mathcal{M} = L\mathcal{K}$)
3. *Form larger* $\mathcal{K}$, *by choosing* $\mathbf{p}_{m+1}$ *linearly independent of*
   $\mathbf{p}_0, .., \mathbf{p}_m$

In Petrov-Galerkin-Krylov methods $\mathcal{K}$ is taken to be the Krylov subspace

$$\mathcal{K}_m(L, \mathbf{r}_0) \equiv span\{\mathbf{r}_0, L\mathbf{r}_0, ..., L^{m-1}\mathbf{r}_0\} \tag{12}$$

3.2 Choice of basis vectors for $\mathcal{K}_m$

In step 3 of the generic PG algorithm $\mathbf{p}_{m+1}$ is guaranteed to be independent of $\mathbf{p}_0$,..,$\mathbf{p}_m$ if either of the following two conditions hold (assuming nonsingular L):

$$\langle \mathbf{p}_{m+1}, \mathbf{p}_j \rangle = 0, \ or \ \langle L\mathbf{p}_{m+1}, L\mathbf{p}_j \rangle = 0, \ j = 0, .., m \tag{13}$$

ie the $\mathbf{p}_i$'s are required, respectively, to be orthogonal (as in GMRES) or L^TL-orthogonal (as in GCR). In the PGK algorithm of Section 3.4 below this choice is made by specifying a bilinear form b on R^n, ie we require

$$b(\mathbf{p}_{m+1}, \mathbf{p}_j) = 0, \ j = 0, .., m \tag{14}$$

where b(·,·)=$\langle \cdot, \cdot \rangle$ or $\langle L\cdot, L\cdot \rangle$. Plainly, (14) does not uniquely determine $\mathbf{p}_{m+1}$. Common choices are listed as 3-i and 3-ii in the RT-PGK algorithm of Section 3.4 below.

3.3 Restarts and truncations

Steps 2 and 3 of the generic PG algorithm listed above imply ever increasing storage and cost per iteration as the number of iterations rises. To save storage and cost PGK methods are usually truncated or restarted. 'Truncation' means only using

$$\mathbf{p}_{\phi(m)}, .., \mathbf{p}_m$$

instead of $\mathbf{p}_0$,..,$\mathbf{p}_m$ for the orthogonalisation in step 2, and in imposing (13) in step 3 only requiring orthogonality (or L^TL-orthogonality) of P_{m+1} against

$$\mathbf{p}_{\theta(m)}, .., \mathbf{p}_m$$

instead of against all $\mathbf{p}_0$,..,$\mathbf{p}_m$, where θ, ϕ are monotonically increasing integer-valued functions of the non-negative integers such that

$$0 \le \theta(i), \phi(i) \le i$$

Using Sections 3.2 and 3.3 in the generic PG method of Section 3.1 gives:

3.4 Restarted/Truncated Petrov-Galerkin-Krylov Algorithm RT-PGK(b, θ, ϕ)

(1) *Given* $\mathbf{x}_0$ *set* $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{f} - L\mathbf{x}_0$, $i = 0$

(2) $\mathbf{r}_{i+1} = \mathbf{r}_i - \sum_{j=\phi(i)}^{i} \alpha_j^i L\mathbf{p}_j$

where α_j^i are chosen so that either

(*i*) $\langle \mathbf{r}_{i+1}, \mathbf{p}_j \rangle = 0$ *for* $\phi(i) \le j \le i$ (*corresponding to choosing* $\mathcal{M} = \mathcal{K}$)

or

(*ii*) $\langle \mathbf{r}_{i+1}, L\mathbf{p}_j \rangle = 0$ *for* $\phi(i) \le j \le i$ (*corresponding to choosing* $\mathcal{M} = L\mathcal{K}$)

(3) Calculate $\mathbf{p}_{i+1}$ either by

(*i*) $\mathbf{p}_{i+1} = \mathbf{r}_{i+1} + \sum_{j=\theta(i)}^{i} \beta_j^i \mathbf{p}_j$

or by

(*ii*) $\beta_{i+1}^i \mathbf{p}_{i+1} = L\mathbf{p}_i + \sum_{j=\theta(i)}^{i} \beta_j^i \mathbf{p}_j$

where β_j^i satisfy

$b(\mathbf{p}_{i+1}, \mathbf{p}_j) = 0$ *for* $\theta(i) \le j \le i$

(in case (ii) either β_{i+1}^i=1 or choose β_{i+1}^i such that $|\mathbf{p}_i|$=1 for all i)

4 Either

(i) Unrestarted RT − PGK : Stop if $\mathbf{r}_i$ is sufficiently small, else
iterate i → i + 1 and go to step 2
or

(ii) Restarted RT − PGK
If i < k : iterate i → i + 1 and go to step 2
If i = k : If $\mathbf{r}_k$ is sufficiently small Stop
else set $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{r}_k$ and go to step 2

3.5 MO GCR, Standard (EES) GCR, GMRES and Orthomin as Special Cases of RT-PGK

(a) Standard (EES) GCR(k) $b(\mathbf{u},\mathbf{v}) \equiv (L\mathbf{u}, L\mathbf{v})$, $\theta(i) = \phi(i) = 0$ *all* i

Unrestarted GCR is RT-PGK(b, θ, ϕ) with formulation 2-(ii), 3-(i), 4-(i), the standard GCR(k) of EES is RT-PGK(b, θ, ϕ) with formulation 2-(ii), 3-(i), 4-(ii).

(b) GMRES(k) $b(\mathbf{u},\mathbf{v}) \equiv (\mathbf{u},\mathbf{v})$, $\theta(i) = \phi(i) = 0$ *all* i

Unrestarted GMRES is RT-PGK(b, θ, ϕ) with formulation 2-(ii), 3-(ii), 4-(i), GMRES(k) is RT-PGK(b, θ, ϕ) with formulation 2-(ii), 3-(ii), 4-(ii).

(c) Orthomin(k) $b(\mathbf{u},\mathbf{v}) \equiv (L\mathbf{u}, L\mathbf{v})$, $\theta(i) = \phi(i) = max\{0, i-k\}$

Orthomin(k) is RT-PGK(b, θ, ϕ) with formulation 2-(ii), 3-(i), 4-(i)

(d) MO GCR(k) $b(\mathbf{u},\mathbf{v}) \equiv (L\mathbf{u}, L\mathbf{v})$, $\theta(i) = \phi(i) = k[i/k]$ ([.] denotes integer part)

MO GCR(k) is RT-PGK(b, θ, ϕ) with formulation 2-(ii), 3-(i), 4-(i)

Thus in the framework of [8] MO GCR(k) and Orthomin(k) are truncated methods while standard (EES) GCR(k) is a restarted method.

4 Link with Damped Wave Equation

GCR algorithms are often derived by discretising a damped wave equation ([10],[9],[11]). [10] show how Orthomin(1) (there designated the conjugate residual method) which is identical to MO GCR(1) may be obtained by discretising the damped wave equation

$$\frac{\partial^2\phi}{\partial\tau^2} + \frac{1}{T(\tau)}\frac{\partial\phi}{\partial\tau} = \mathcal{L}(\phi) - R$$

whose steady state solution in pseudo-time τ represents the solution to the governing problem

$$\mathcal{L}(\phi) = R$$

It has been stated in the literature that a similar discretisation of

$$\frac{\partial^k\phi}{\partial\tau^k} + \frac{1}{T_{k-1}(\tau)}\frac{\partial^{k-1}\phi}{\partial\tau^{k-1}} + ... + \frac{1}{T_1(\tau)}\frac{\partial\phi}{\partial\tau} = \mathcal{L}(\phi) - R \qquad (15)$$

for any k leads to MO GCR(k-1). In fact discretising (15) yields Orthomin(k-1). This is demonstrated for k=3 in Appendix A, where it is shown that discretising (15) with k=3 gives the algorithm:

Given $\mathbf{r}_n$, $\mathbf{p}_n$, $\mathbf{p}_{n-1}$

Calculate β_n by (101)
Calculate $\mathbf{r}_{n+1}$ by (94)
Calculate A_{n+1} and B_{n+1} by (102,103)
Calculate $\mathbf{p}_{n+1}$ by (93)

which is Orthomin(2). It differs from MO GCR(2) essentially in eqn 93. MO GCR(2) is as follows:

Given $\mathbf{r}_n$, $\mathbf{p}_n$

Calculate β_n by (101)
Calculate $\mathbf{r}_{n+1}$ by (94)
Calculate A_{n+1} by (102)
Calculate $\mathbf{p}_{n+1} = \mathbf{r}_{n+1} + A_{n+1}\mathbf{p}_n$
———————————
Calculate β_{n+1} by (101)
Calculate $\mathbf{r}_{n+2}$ by (94)
Calculate A_{n+2} and B_{n+2} by (102,103)
Calculate $\mathbf{p}_{n+2}$ by (93)

Note that the first half of MO GCR(2) is an Orthomin(1) iteration and the second half an Orthomin(2) iteration. In general, as mentioned in Section 2, each cycle of k MO GCR(k) iterations consists of an Orthomin(1) iteration followed by an Orthomin(2) iteration followed by ... followed by an Orthomin(k) iteration.

It is only in this sense that discretising (15) yields MO GCR(k-1).

5 Properties of The Iterates $\mathbf{r}_i$, $\mathbf{p}_i$ in MO GCR(k)

We now establish the fundamental properties of the MO GCR(k). In order to keep track of the quantities generated by the algorithm over more than one loop it is convenient to formulate MO GCR(k) as follows (cf formulation in §2) which also allows for iteration dependent preconditioner P_i^{-1}[2]:

Given an initial guess $\mathbf{x}_0$, set

$$\mathbf{r}_0 = \mathbf{f} - L\mathbf{x}_0 \; and \; \mathbf{p}_0 = P_0^{-1}\mathbf{r}_0 \tag{16}$$

For m=0 until converged do:
For i=mk+1 to mk+k do:
Calculate in order a_{i-1}, $\mathbf{x}_i$, $\mathbf{r}_i$,
$(b_{mk}^{i-1}, b_{mk+1}^{i-1}, \ldots, b_{i-1}^{i-1})$, $\mathbf{p}_i$
End do
End do

[2]Conventionally preconditioned algorithms for CG-type methods are obtained by writing down the unpreconditioned algorithm for $\mathrm{LP}^{-1}\mathrm{P}\mathbf{x}=\mathbf{f}$ and re-arranging. This gives the algorithm listed if $\mathrm{P}_i=\mathrm{P}$ constant, but we need to allow for P_i varying with iteration number. In the Met Office Helmholtz solver applications $\mathrm{P}_0 \neq \mathrm{P}_1 = \mathrm{P}_j$ for all $j \geq 1$.

where a_i, $\mathbf{x}_{i+1}$, $\mathbf{r}_{i+1}$ are as defined in (2-4) and for i=mk,mk+1,...,(m+1)k-1

$$\mathbf{p}_{i+1} = P_{i+1}^{-1}\mathbf{r}_{i+1} + b_{mk}^i \mathbf{p}_{mk} + b_{mk+1}^i \mathbf{p}_{mk+1} + ... + b_i^i \mathbf{p}_i \tag{17}$$

where

$$b_j^i = -\frac{\langle LP_{i+1}^{-1}\mathbf{r}_{i+1}, L\mathbf{p}_j\rangle}{|L\mathbf{p}_j|^2} \tag{18}$$

Henceforth 'MO GCR(k)' refers to this preconditioned algorithm unless otherwise specified. We may readily check that (8) still holds so if L is invertible $\mathbf{x}_i$ may be deduced from $\mathbf{r}_i$, so the algorithm is analysed in terms of $\{\mathbf{r}_i,\mathbf{p}_i\}$. Other properties of this algorithm are established below and in Section 6.

5.1 Relations between the Iterates in The MO GCR(k)

The relations given in Section 3 of [2] are for their (ie 'EES') GCR(k) and for L having positive definite symmetric part, ie for

$$L_S \equiv \frac{1}{2}(L + L^T)$$

positive definite[3]. For example, for EES GCR(k-1)

$$\begin{aligned} span(\mathbf{p}_{mk}, \mathbf{p}_{mk+1}, ..., \mathbf{p}_i) = \\ span(\mathbf{p}_{mk}, L\mathbf{p}_{mk}, ..., L^{i-mk}\mathbf{p}_{mk}) = \\ span(\mathbf{r}_{mk}, \mathbf{r}_{mk+1}, ..., \mathbf{r}_i) \end{aligned}$$

for any $mk \le i \le (m+1)k$ but this is not true for the MO GCR(k) because in this case each loop does not begin with a complete restart (note (1) of Section 2). Other relations established in [2] do still hold for MO GCR(k) if L_S is positive definite, eg $\mathbf{r}_i \neq 0$ implies $\mathbf{p}_i \neq 0$ (see footnote in Section 5.2), but this is not true for general L. However, most of the relations established in [2] for GCR (ie EES GCR(∞), which is the same as MO GCR(∞)) do still hold for MO GCR(k) if applied over the appropriate range[4]. For all m $\ge$ 0:

$$\langle L\mathbf{p}_i, L\mathbf{p}_j\rangle = 0 \text{ any } mk \le i \ne j \le (m+1)k \tag{19}$$

$$\langle \mathbf{r}_i, L\mathbf{p}_j\rangle = 0 \text{ any } mk \le j < i \le (m+1)k \tag{20}$$

for all i $\ge$ 0:

$$\langle \mathbf{r}_i, L\mathbf{p}_i\rangle = \langle \mathbf{r}_i, LP_i^{-1}\mathbf{r}_i\rangle \tag{21}$$

For all m $\ge$ 0:

$$\langle \mathbf{r}_i, LP_j^{-1}\mathbf{r}_j\rangle = 0 \text{ any } mk \le j < i \le (m+1)k \tag{22}$$

$$\langle \mathbf{r}_j, L\mathbf{p}_i\rangle = \langle \mathbf{r}_{mk}, L\mathbf{p}_i\rangle \text{ any } mk \le j \le i \le (m+1)k \tag{23}$$

for all i $\ge$ 0:

$$\langle L\mathbf{p}_i, L\mathbf{p}_i\rangle = \langle L\mathbf{p}_i, LP_i^{-1}\mathbf{r}_i\rangle \tag{24}$$

$$|L\mathbf{p}_i|^2 \le |LP_i^{-1}\mathbf{r}_i|^2 \tag{25}$$

---

[3]This implies that L is positive stable, ie has all its eigenvalues lying entirely in the open right half-plane. The converse is not true: there exist positive stable matrices the symmetric parts of which are indefinite, eg

$$\begin{pmatrix} 2 & 4 \\ -4 & -1 \end{pmatrix}.$$

Needless to say, the analysis of [2] applies equally to matrices with negative definite symmetric parts.

[4]Proofs of (19)-(25) are similar to those established in Section 3 of [2] for their relations and are given in Appendix B

5.2 Failure of MO GCR(k)

Whether the sequence $\mathbf{x}_i$ generated by the MO GCR(k) converges to $\mathrm{L}^{-1}\mathbf{f}$ or not, scrutiny of the algorithm above and in Section 2 shows that the algorithm itself can only fail at the ith iterate (fail in the sense of being unable to generate an (i+1)th iterate) if $\mathrm{L}\mathbf{p}_i=\mathbf{0}$. It follows from the analysis of [2] that this can never happen for their GCR(k) if the symmetric part of L is definite, and this is also true for MO GCR(k) if LP_i^{-1} is definite[5]. For general L (including L's with indefinite symmetric part) it is evident that the condition

$$\mathrm{L}\mathbf{p}_i=\mathbf{0} \text{ for some i}$$

can only occur for a subset of measure zero in the n^2+2n space of $n\times n$ matrices L, n-vectors $\mathbf{f}$ and n-vectors $\mathbf{x}_0$ representing all possible equations and initial states of the algorithm.

5.3 Convergence of the Modulii of the Residuals of MO GCR(k)

It is clear from (4,3) that for any L the magnitude of the residual $|\mathbf{r}_i|$ decreases with i:

$$\begin{aligned}|\mathbf{r}_{i+1}|^2 &= |\mathbf{r}_i|^2 + a_i^2|L\mathbf{p}_i|^2 - 2a_i\langle\mathbf{r}_i, L\mathbf{p}_i\rangle \\ &= |\mathbf{r}_i|^2 + \frac{\langle\mathbf{r}_i, L\mathbf{p}_i\rangle^2}{|L\mathbf{p}_i|^2} - 2\frac{\langle\mathbf{r}_i, L\mathbf{p}_i\rangle^2}{|L\mathbf{p}_i|^2} \le |\mathbf{r}_i|^2\end{aligned} \tag{26}$$

Since $|\mathbf{r}_i|$ is decreasing in i it either reaches 0 or converges to a positive scalar.

5.4 Convergence of the Difference of the Residuals of MO GCR(k)

If the symmetric part of LP_i^{-1} is indefinite for arbitrarily large i there is no guarantee that the sequence $\mathbf{x}_i$ generated by the MO GCR(k) will converge to $L^{-1}\mathbf{f}$, or equivalently that $\mathbf{r}_i$ will converge to $\mathbf{0}$ (see §6.1 below). This leaves open the possibility that these sequences could oscillate for ever[6], or display more complex long-term behaviour. The following result limits the long-term behaviour:

Proposition 1

For any matrix L the sequence $\mathbf{r}_i$ generated by the MO GCR(k) satisfies

$$(\mathbf{r}_{i+1}-\mathbf{r}_i)^2 = |\mathbf{r}_i|^2 - |\mathbf{r}_{i+1}|^2$$

and hence by (5.3)

$$\mathbf{r}_{i+1}-\mathbf{r}_i \to \mathbf{0}$$

as i $\to\infty$.

Proof See Appendix B.

5.5 Homogeneity of MO GCR(k)

Suppose we start the MO GCR(k) from an initial residual

$$\hat{\mathbf{r}}_0 = \lambda\mathbf{r}_0$$

for some scalar λ, and hence by definition the corresponding

$$\hat{\mathbf{p}}_0 = \lambda\mathbf{p}_0$$

Then from (3,4,17,18) it readily follows by induction that for every i

$$\hat{\mathbf{r}}_i = \lambda\mathbf{r}_i,\ \hat{\mathbf{p}}_i = \lambda\mathbf{p}_i,\ \hat{a}_i = a_i,\ \hat{b}^i_j = b^i_j$$

[5]For the same reason as in [2], ie that if $\mathrm{L}\mathbf{p}_i=\mathbf{0}$ then $\langle\mathbf{r}_i, L\mathbf{p}_i\rangle=\mathbf{0}$ so by (21) $\langle\mathbf{r}_i, LP_i^{-1}\mathbf{r}_i\rangle=0$ contrary to positive definiteness of $(LP_i^{-1})_S$.

[6]Since convergence of $|\mathbf{r}_i|$ as guaranteed by (5.3) does not imply convergence of $\mathbf{r}_i$.

5.6 Stationary State of MO GCR(k)

If for some m $|\mathbf{r}_{mk+1}| = |\mathbf{r}_{mk}|$, and P_j=P_{mk} for all j≥mk, then $\mathbf{r}_j = \mathbf{r}_{mk}$ for all j≥mk (cf Proposition 1).

Proof See Appendix B.

5.7 The Asymptotic Behaviour of the MO GCR(k)

(a) As i $\to \infty$

$$\frac{\langle \mathbf{r}_i, L\mathbf{p}_i \rangle}{|L\mathbf{p}_i|} \to 0 \tag{27}$$

$$\frac{\langle \mathbf{r}_i, LP_i^{-1}\mathbf{r}_i \rangle}{|L\mathbf{p}_i|} \to 0 \tag{28}$$

and hence

$$\langle \mathbf{r}_i, L\mathbf{p}_i \rangle \to 0 \tag{29}$$

$$\langle \mathbf{r}_i, LP_i^{-1}\mathbf{r}_i \rangle \to 0 \tag{30}$$

(b) If for all sufficiently large j $|L\mathbf{p}_j|$ is bounded below by a positive scalar and P_j=P a constant then as m $\to \infty$

$$b_{mk+j}^{mk+i} \to -1 \tag{31}$$

for every $0 \leq j \leq i \leq$ k-1.

Proof See Appendix B.

Corollary 1

If k=1 and for all sufficiently large j $|L\mathbf{p}_j|$ is bounded below by a positive scalar and P_j=P a constant then

$$\mathbf{p}_{m+2} - \mathbf{p}_m \to 0 \tag{32}$$

$$\mathbf{p}_{m+1} + \mathbf{p}_m - P^{-1}\mathbf{r}_{m+1} \to 0 \tag{33}$$

as m $\to \infty$

Proof

(17) and (31) give (33) and

$$\mathbf{p}_{m+2} + \mathbf{p}_{m+1} - P^{-1}\mathbf{r}_{m+2} \to \mathbf{0} \tag{34}$$

and subtracting (33) from (34) and using Proposition 1 gives (32).

Remark

The requirement of 5.7(b) and Corollary 1 that $|L\mathbf{p}_i|$ does not tend to 0 can, in view of (25), only occur in cases of non-convergence ($\mathbf{r}_i$ does not tend to $\mathbf{0}$). In this case the $\mathbf{p}_i$'s cannot converge (for if they did (33) would imply $2\mathbf{p}_i - P^{-1}\mathbf{r}_i \to \mathbf{0}$, and (18) would imply $b_j^i \to$-1/2 contrary to (31)). Thus if the conditions of Corollary 1 hold the simplest behaviour of the $\mathbf{p}_i$'s allowed by (32) is to oscillate between two states – see Section 7.1 below.

6 Conditions for Convergence of The MO GCR(k) Algorithm

At the beginning of each loop of EES's GCR(k) $\mathbf{p}_{mk}$ is reset to $\mathbf{r}_{mk}$, and this leads to the particularly simple characterisation (35) of $|\mathbf{r}_{mk+i}|$ for that method. For MO GCR(k) there is no such simple characterisation, but a weak version of (35) applicable to MO GCR(k) (Proposition 2) is used to give conditions for convergence of the algorithm to the required solution (Corollory 2). In particular, as with EES's GCR(k), definiteness of the symmetric part of LP_i^{-1} for all sufficiently large i is a sufficient condition for convergence.

6.1 Minimisation Property and Convergence of MO GCR(k)

The standard GCR algorithm of [2] listed in Section 2 has a well-known minimisation characterisation, which we may state as follows: if i $\leq$ k then $\mathbf{r}_{mk+i}$ as generated by EES GCR(k-1) is the vector of minimum norm in

$$\mathbf{r}_{mk} + span\{L\mathbf{r}_{mk}, L^2\mathbf{r}_{mk}..., L^i\mathbf{r}_{mk}\}$$

Hence if we denote the set of monic ith degree polynomials[7] by P_i then the residuals $\mathbf{r}_{mk+i}$ generated by EES GCR(k-1) are of the form

$$\mathbf{r}_{mk+i} = q_i(L)\mathbf{r}_{mk} \qquad EES\,GCR(k-1)\ only$$

for some $q_i \in P_i$ and

$$|\mathbf{r}_{mk+i}| = \min_{q_i \in P_i} |q_i(L)\mathbf{r}_{mk}| \qquad EES\,GCR(k-1)\ only \tag{35}$$

None of this is true for MO GCR(k), the closest we come is in Proposition 2 below.

Applying (3) recursively we see that

$$\mathbf{r}_{mk+i} = \mathbf{r}_{mk} - a_{mk}L\mathbf{p}_{mk} - ... - a_{mk+i-1}L\mathbf{p}_{mk+i-1} \tag{36}$$

so using (19)

$$|\mathbf{r}_{mk+i}|^2 = |\mathbf{r}_{mk}|^2 + |a_{mk}L\mathbf{p}_{mk}|^2 + ... + |a_{mk+i-1}L\mathbf{p}_{mk+i-1}|^2 - 2\langle \mathbf{r}_{mk}, a_{mk}L\mathbf{p}_{mk} + ... + a_{mk+i-1}L\mathbf{p}_{mk+i-1}\rangle \tag{37}$$

If we were to minimise (37) over a_{mk},...,a_{mk+i-1} we would get using (23) that

$$2a_j|L\mathbf{p}_j|^2 - 2\langle \mathbf{r}_j, L\mathbf{p}_j\rangle = 0$$

each j=mk,...,mk+i-1. These a_j are exactly as given by (4), ie, a_{mk},...,a_{mk+i-1} as determined by (4) minimise

$$|\mathbf{r}_{mk} - a_{mk}L\mathbf{p}_{mk} - ... - a_{mk+i-1}L\mathbf{p}_{mk+i-1}|$$

Thus, for any i $\leq$ k:

$$\mathbf{r}_{mk+i} \text{ is the vector of minimum norm in } \mathbf{r}_{mk} + span\{L\mathbf{p}_{mk}, L\mathbf{p}_{mk+1}..., L\mathbf{p}_{mk+i-1}\} \tag{38}$$

In words, at the ith stage of the mth loop MO GCR(k) and EES GCR(k-1) minimise the error $\mathbf{r}_{mk+i}$ over the affine space (38). Equivalently, for any i $\leq$ k

$$\mathbf{x}_{mk+i} \text{ minimises } |\mathbf{f} - L\mathbf{x}_{mk+i}| \text{ over } \mathbf{x}_{mk} + span\{\mathbf{p}_{mk}, ..., \mathbf{p}_{mk+i-1}\} \tag{39}$$

Because in EES GCR(k-1) each loop begins with a genuine restart, ie

$$\mathbf{r}_{mk} = \mathbf{p}_{mk} \qquad EES\,GCR(k-1)\ only \tag{40}$$

[7] ie, polynomials of the form $1+\alpha_1 x + \alpha_2 x^2 + ... + \alpha_i x^i$

we have using (3), (17) (with P_{i+1}=I) and (40) that

$$span\{L\mathbf{p}_{mk}, L\mathbf{p}_{mk+1}..., L\mathbf{p}_{mk+i-1}\} = span\{L\mathbf{r}_{mk}, L^2\mathbf{r}_{mk}..., L^i\mathbf{r}_{mk}\} \qquad EES\ only \tag{41}$$

(the Krylov subspace based on L, $\mathbf{r}_{mk}$) and the characterisation of $\mathbf{r}_{mk+i}$ as generated by EES GCR(k-1) at the beginning of this section follows. It is not true for MO GCR(k) because (40) and hence (41) is not true for MO GCR(k). However

Proposition 2

The residuals generated by MO GCR(k) have the property that for any k $\geq$ 1

$$|\mathbf{r}_{mk+1}| \leq \min_{q_1 \in P_1} |q_1(LP_{mk}^{-1})\mathbf{r}_{mk}| \tag{42}$$

Proof By (38) for any α

$$|\mathbf{r}_{mk+1}|^2 \leq |\mathbf{r}_{mk} + \alpha L\mathbf{p}_{mk}|^2 \tag{43}$$

By (21) and (25) for any α

$$\begin{aligned} |\mathbf{r}_{mk}|^2 + \alpha^2|L\mathbf{p}_{mk}|^2 + 2\alpha\langle\mathbf{r}_{mk}, L\mathbf{p}_{mk}\rangle \leq \\ |\mathbf{r}_{mk}|^2 + \alpha^2|LP_{mk}^{-1}\mathbf{r}_{mk}|^2 + 2\alpha\langle\mathbf{r}_{mk}, LP_{mk}^{-1}\mathbf{r}_{mk}\rangle \end{aligned} \tag{44}$$

and so by (43) and (44) for any α

$$|\mathbf{r}_{mk+1}|^2 \leq |\mathbf{r}_{mk} + \alpha LP_{mk}^{-1}\mathbf{r}_{mk}|^2$$

as claimed.

A comparison of (42) applicable to MO GCR(1) and (35) applicable to EES GCR(k-1) suggests that ND GCR(1) converges at least as fast as EES GCR(0), which [2] show does converge if L has positive definite symmetric part. This is formalised in Corollary 2 which relies on the following fact which [2] establish during the course of proving their Theorem 3.3: that if M_S, the symmetric part of matrix M, is positive definite then

$$\left\| I - \frac{\lambda_{min}(M_S)}{\lambda_{max}(M^TM)} M \right\|_2^2 \leq 1 - \frac{[\lambda_{min}(M_S)]^2}{\lambda_{max}(M^TM)} \tag{45}$$

where

$$\|M\|_2 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|M\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Corollary 2

If there exists a positive integer i_0 and a positive scalar δ with

$$\lambda_{min}[(LP_i^{-1})_S] \geq \delta \tag{46}$$

for all i$\geq i_0$ then MO GCR(k) converges to the required solution for any k $\geq$ 1

Proof By Proposition 2 for any k $\geq$ 1

$$|\mathbf{r}_{mk+k}| \leq \min_{\alpha\in R} |\mathbf{r}_{mk} + \alpha LP_{mk}^{-1}\mathbf{r}_{mk}| \leq \min_{\alpha\in R} \|I + \alpha LP_{mk}^{-1}\|_2 |\mathbf{r}_{mk}|$$

Choosing

$$\alpha = -\frac{\lambda_{min}((LP_{mk}^{-1})_S)}{\lambda_{max}((LP_{mk}^{-1})^T(LP_{mk}^{-1})L)}$$

the result follows by (45) and (46).

To maximise convergence rate (see Section 8.3) P_i^{-1} are chosen so that the eigenvalues of LP_i^{-1} lie in the right half plane and as close to 1 as possible, so condition (46) should normally be met.

6.2 Convergence of MO GCR(k) For Large k

From the characterisation (39) of the iterates $\mathbf{x}_{i+1}$ generated by MO GCR(k) it is clear that the larger the value of k the larger the subspace over which we minimise the error at each stage and hence both the faster the convergence and the better the chance of converging at all. If the algorithm doesn't fail then in the limit as $k \to \infty$ $\mathbf{x}_i$ will converge to $L^{-1}\mathbf{f}$ (this is true for both EES and MO GCR(k)). The proof is similar to that given in [2] for the definite case. By (39) if $i+1 \leq k$ the $\mathbf{x}_{i+1}$ chosen by GCR(k) has the property that it minimises $|L\,\mathbf{x} - \mathbf{f}|$ over

$$\mathbf{x}_0 + span\{\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_i\}$$

The subspace $span\{\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_i\}$ will eventually (ie for large enough i) include the origin (and hence $|L\,\mathbf{x} - \mathbf{f}|=0$) if the vectors

$$\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_i$$

are independent. By assumption the vectors

$$L\mathbf{p}_0, L\mathbf{p}_1, ..., L\mathbf{p}_i$$

are non-zero so they must be independent (since if

$$\lambda_0 L\mathbf{p}_0 + \lambda_1 L\mathbf{p}_1 + \lambda_i L\mathbf{p}_i = \mathbf{0}$$

then taking the inner product of this expression with $L\mathbf{p}_j$ for j=0,...i and using (19) shows that $\lambda_j=0$ for all j) and so for nonsingular L $\{\mathbf{p}_j$, j=0,...,i$\}$ are independent and the result follows.

(5.2) and (6.2) above mean that the MO GCR(k) almost never fails, and if it doesn't fail will converge to the required solution for large enough k.

7 Geometric perspective: MO GCR(k) as an Iterated Map

For simplicity here we consider only the unpreconditioned case with P_i=I for all i. Given an initial guess $\mathbf{x}_0$ MO GCR(k) generates the sequence

$$\mathbf{x}_0 \to \mathbf{x}_1 \to \mathbf{x}_2 \to ...$$

$\mathbf{x}_1$ is a function of $\mathbf{x}_0$ but for i>1 $\mathbf{x}_i$ is a function of all of $\mathbf{x}_0$,...,$\mathbf{x}_{i-1}$; to view GCR(k) as an iterated map we want to put it in the form

$$\mathbf{z}' = G\mathbf{z}$$

This can be done in more than one way. The simplest is as an iterated map on R^{2kn}. In this case the state variable is

$$\mathbf{z} = (\mathbf{r}_1,\ \mathbf{p}_1,\ \mathbf{r}_2,\ \mathbf{p}_2,\ \ldots,\ \mathbf{r}_k,\ \mathbf{p}_k)^T \in R^{2kn}$$

From the MO GCR(k) algorithm as given in Section 2 we may form

$$\mathbf{z}' = \begin{pmatrix} \mathbf{r}_1'(\mathbf{r}_k, \mathbf{p}_k) \\ \mathbf{p}_1'(\mathbf{p}_k, \mathbf{r}_1') \\ \mathbf{r}_2'(\mathbf{r}_1', \mathbf{p}_1') \\ \mathbf{p}_2'(\mathbf{p}_k, \mathbf{p}_1', \mathbf{r}_2') \\ \cdot \\ \cdot \\ \mathbf{r}_k'(\mathbf{r}_{k-1}', \mathbf{p}_{k-1}') \\ \mathbf{p}_k'(\mathbf{p}_k, \mathbf{p}_1', .., \mathbf{p}_{k-1}', \mathbf{r}_k') \end{pmatrix}$$

Alternatively at the cost of extra complexity we may express MO GCR(k) as an iterated map on $R^{(k+1)n}$ (this is the smallest space possible). In this case the state variable is

$$\mathbf{z} = (\mathbf{r}_1,\ \mathbf{p}_1,\ \mathbf{p}_2,\ \ldots, \mathbf{p}_k)^T \in R^{(k+1)n}$$

From the MO GCR(k) algorithm as given in Section 2 we may form

$$\mathbf{z}' = \begin{pmatrix} \mathbf{r}'_1(\mathbf{r}_k(\mathbf{r}_{k-1}(...\mathbf{r}_4(\mathbf{r}_3(\mathbf{r}_2(\mathbf{r}_1,\mathbf{p}_1),\mathbf{p}_2),\mathbf{p}_3),...),\mathbf{p}_{k-1}),\mathbf{p}_k) \\ \mathbf{p}'_1(\mathbf{p}_k,\mathbf{r}'_1) \\ \mathbf{p}'_2(\mathbf{p}_k,\mathbf{p}'_1,\mathbf{r}'_2(\mathbf{r}'_1,\mathbf{p}'_1)) \\ \mathbf{p}'_3(\mathbf{p}_k,\mathbf{p}'_1,\mathbf{p}'_2,\mathbf{r}'_3(\mathbf{r}'_2(\mathbf{r}'_1,\mathbf{p}'_1),\mathbf{p}'_2)) \\ . \\ . \\ \mathbf{p}'_k(\mathbf{p}_k,\mathbf{p}'_1,..,\mathbf{p}'_{k-1},\mathbf{r}'_k(\mathbf{r}'_{k-1}(...\mathbf{r}'_3(\mathbf{r}'_2(\mathbf{r}'_1,\mathbf{p}'_1),\mathbf{p}'_2),...),\mathbf{p}'_{k-1})) \end{pmatrix}$$

Evidently these maps are nonlinear, but as shown in Section 5.5 are homogeneous in the sense that G(λ **z**)=λG(**z**). If L is definite then by Corollary 2 $\mathbf{r}_i \to \mathbf{0}$ and by (25) $\mathbf{p}_i \to \mathbf{0}$. Hence in this case **0** is a global attractor for G: for any initial $\mathbf{r}_0$ $G^i(\mathbf{z}_0) \to \mathbf{0}$ as i$\to \infty$. From the geometric dynamics point of view the more intersting case is when the algorithm *doesn't* converge to the required solution.

7.1 Example: MO GCR(1)

For convenience we shall concentrate on the MO GCR(1) as the paradigm for all MO GCR(k)'s. This is the same as [2]'s orthomin(1), though the analysis of it in [2] is for the case where L has positive definite symmetric part (as shown above, in this case **0** is a global attractor). From Section 2 we see that the MO GCR(1) may be formulated as the iterated map

$$\begin{pmatrix} \mathbf{r} \\ \mathbf{p} \end{pmatrix} \to \begin{pmatrix} \mathbf{r}' \\ \mathbf{p}' \end{pmatrix}$$

where

$$\mathbf{r}' = \mathbf{r} - a(\mathbf{r},\mathbf{p})L\mathbf{p} \tag{47}$$

$$\mathbf{p}' = \mathbf{r}' + b(\mathbf{r}',\mathbf{p})\mathbf{p} \tag{48}$$

where

$$a(\mathbf{r},\mathbf{p}) = \frac{\langle \mathbf{r}, L\mathbf{p}\rangle}{\langle L\mathbf{p}, L\mathbf{p}\rangle} \tag{49}$$

$$b(\mathbf{r}',\mathbf{p}) = -\frac{\langle L\mathbf{r}', L\mathbf{p}\rangle}{\langle L\mathbf{p}, L\mathbf{p}\rangle} \tag{50}$$

Equivalently

$$\begin{pmatrix} \mathbf{r} \\ \mathbf{p} \end{pmatrix} \to \begin{pmatrix} \mathbf{r}' \\ \mathbf{p}' \end{pmatrix} = \begin{pmatrix} I - \frac{L\mathbf{p}\mathbf{p}^T L^T}{|L\mathbf{p}|^2} \\ \\ \left(I - \frac{\mathbf{p}\mathbf{p}^T L^T L}{|L\mathbf{p}|^2}\right)\left(I - \frac{L\mathbf{p}\mathbf{p}^T L^T}{|L\mathbf{p}|^2}\right) \end{pmatrix} \mathbf{r} \tag{51}$$

Evidently we are particularly interested in the first component of (51) since if $\mathbf{r}_i$ does not converge to **0** the algorithm has not converged to the true solution. It was shown in Section 5.3 that $|\mathbf{r}_i|$ always converges and in section 5.4 that $\mathbf{r}_{i+1} - \mathbf{r}_i$ always converges to **0**. Let us now consider non-zero fixed points of the first component of (51). Let $\mathbf{r}_b$, $\mathbf{p}_b$ be such that

$$L\mathbf{p}_b \neq \mathbf{0},\ L(\mathbf{r}_b - \mathbf{p}_b) \neq \mathbf{0}\,. \tag{52}$$

Note that (52) implies $\mathbf{r}_b \neq \mathbf{0}$, since otherwise if $\mathbf{r}_b$=**0** then by (25) L$\mathbf{p}_b$=**0** so (52) cannot be satisfied. We now show that MO GCR(1) leaves $\mathbf{r}_b$ fixed if and only if the following two conditions hold:

$$\langle \mathbf{r}_b, L\mathbf{p}_b\rangle = 0 \tag{53}$$

$$\langle \mathbf{r}_b, L\mathbf{r}_b\rangle = 0 \tag{54}$$

and in this case the component $\mathbf{p}_b$ has period 2

$$\mathbf{p}_b \to \mathbf{r}_b - \mathbf{p}_b \to \mathbf{p}_b\,. \tag{55}$$

Proof 'if': (52,53) imply $a(\mathbf{r}_b, \mathbf{p}_b) = 0$ so $\mathbf{r}'_b = \mathbf{r}_b$. This fact, (52) and (24) imply that $b(\mathbf{r}'_b, \mathbf{p}_b) = -1$ so $\mathbf{p}'_b = \mathbf{r}_b - \mathbf{p}_b$. By (52,53,54)

$$a(\mathbf{r}_b, \mathbf{p}'_b) = \frac{\langle \mathbf{r}_b, L(\mathbf{r}_b - \mathbf{p}_b)\rangle}{|L(\mathbf{r}_b - \mathbf{p}_b)|^2} = 0\,,$$

so

$$\mathbf{r}''_b = \mathbf{r}_b$$

$$b(\mathbf{r}''_b, \mathbf{p}'_b) = -\frac{\langle L\mathbf{r}_b, L(\mathbf{r}_b - \mathbf{p}_b)\rangle}{|L(\mathbf{r}_b - \mathbf{p}_b)|^2} = -\frac{|L\mathbf{r}_b|^2 - |L\mathbf{p}_b|^2}{|L\mathbf{r}_b|^2 - |L\mathbf{p}_b|^2}$$

so

$$\mathbf{p}''_b = \mathbf{r}''_b - \mathbf{p}'_b = \mathbf{p}_b$$

'only if': if (52) holds and $\mathbf{r}'_b = \mathbf{r}_b$ then $a(\mathbf{r}_b, \mathbf{p}_b) = 0$ so $\langle \mathbf{r}_b, L\mathbf{p}_b\rangle = 0$. By (21) $\langle \mathbf{r}_i, L\mathbf{p}_i\rangle = \langle \mathbf{r}_i, L\mathbf{r}_i\rangle$ hence $\langle \mathbf{r}_b, L\mathbf{r}_b\rangle = 0$ which concludes the proof.

This means that if the first component of (51) is fixed under MO GCR(1) its second component has period 2. If we represent the maps

$$\begin{pmatrix} \mathbf{r} \\ \mathbf{p} \end{pmatrix} \to \begin{pmatrix} \mathbf{r}' \\ \mathbf{p}' \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}' \\ \mathbf{p}' \end{pmatrix} \to \begin{pmatrix} \mathbf{r}'' \\ \mathbf{p}'' \end{pmatrix}$$

by G and G′ respectively, so that for any positive integers i,m

$$\begin{pmatrix} \mathbf{r}_{i+2m} \\ \mathbf{p}_{i+2m} \end{pmatrix} = (G'G)^m \begin{pmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{pmatrix},$$

then in view of the above result any point

$$\begin{pmatrix} \mathbf{r}_b \\ \mathbf{p}_b \end{pmatrix}$$

in R^{2n} whose first component $\mathbf{r}_b$ is fixed under GCR(1) is a fixed point of G′G. Such a fixed point can in practice only represent a final state of MO GCR(1) if it is an attractor of G′G, since otherwise if

$$\begin{pmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{pmatrix} \text{ is close to } \begin{pmatrix} \mathbf{r}_b \\ \mathbf{p}_b \end{pmatrix}$$

for some i then

$$(G'G)^m \begin{pmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{pmatrix}$$

will not remain close to it. Thus a point satisfying (52-54) can represent a final state of MO GCR(1) if all the eigenvalues of the linearisation of G′G about this point are less than or equal to 1, and not otherwise. If we represent this linearisation by M so that

$$G'G \begin{pmatrix} \mathbf{r}_b + \mathbf{r}_e \\ \mathbf{p}_b + \mathbf{p}_e \end{pmatrix} = \begin{pmatrix} \mathbf{r}_b \\ \mathbf{p}_b \end{pmatrix} + M \begin{pmatrix} \mathbf{r}_e \\ \mathbf{p}_e \end{pmatrix} + O(|\mathbf{r}_e|^2 + |\mathbf{p}_e|^2)$$

then it is possible to show the following:

Proposition 3

If (52-54) hold the eigenvalues of M are 1 repeated 2n-4 times, -1 twice, 0 and λ_0 where

$$\lambda_0 = \frac{\mathbf{r}_b.L^2(\mathbf{r}_b - \mathbf{p}_b)\mathbf{r}_b.L^2\mathbf{p}_b}{|L(\mathbf{r}_b - \mathbf{p}_b)|^2|L\mathbf{p}_b|^2} \tag{56}$$

Corollary 3

A point

$$\begin{pmatrix} \mathbf{r}_b \\ \mathbf{p}_b \end{pmatrix}$$

satisying (52) can only be a final state of MO GCR(1) if (53-54) hold and $|\lambda_0| \leq 1$.

Remark If L is symmetric it is readily shown that (52-54) and (24) imply $|\lambda_0| = 1$. If L is not symmetric (52-54) and (24) can hold without $|\lambda_0| \leq 1$, eg

$$L = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix}, \ \mathbf{r}_b = \begin{pmatrix} 2.0 \\ -0.60555 \\ 2.60555 \end{pmatrix}, \ \mathbf{p}_b = \begin{pmatrix} 1.57343 \\ 1.64373 \\ 1.65002 \end{pmatrix}$$

satisfy (52-54) and (24) but λ_0=-1.48361.

8 Rate of Convergence of MO GCR(k) and Choice of Preconditioner

8.1 Relations between the iterates if LP^{-1} is symmetric

Suppose the preconditioner P_i=P a constant. We show that if LP^{-1} is symmetric then for all mk≤j<i≤(m+1)k

$$b_j^i = 0 \tag{57}$$

If mk≤i<(m+1)k then by (3) and (17)

$$L\mathbf{p}_i \in \{\mathbf{r}_i, \mathbf{r}_{i+1}\}$$

and

$$P^{-1}\mathbf{r}_{i+1} \in \{\mathbf{p}_{mk}, .., \mathbf{p}_{i+1}\}$$

hence

$$P^{-1}L\mathbf{p}_i \in \{\mathbf{p}_{mk}, .., \mathbf{p}_{i+1}\} \tag{58}$$

It follows from (58),the definition of b_j^i and symmetry of LP^{-1} that there exist γ_{mk},...,γ_{j+1} such that

$$b_j^i \equiv -\frac{\langle LP^{-1}\mathbf{r}_{i+1}, L\mathbf{p}_j\rangle}{|L\mathbf{p}_j|^2} = -\frac{\langle \mathbf{r}_{i+1}, LP^{-1}L\mathbf{p}_j\rangle}{|L\mathbf{p}_j|^2} = -\frac{1}{|L\mathbf{p}_j|^2}\langle \mathbf{r}_{i+1}, L(\gamma_{mk}\mathbf{p}_{mk} + ... + \gamma_{j+1}\mathbf{p}_{j+1})\rangle \tag{59}$$

By (20) the right hand side of (59) is zero for all mk≤j<i≤(m+1)k which implies (57). In view of (57) equation (17) takes the form

$$\mathbf{p}_{i+1} = P^{-1}\mathbf{r}_{i+1} + b_i^i\mathbf{p}_i \tag{60}$$

where

$$b_i^i = -\frac{\langle LP^{-1}\mathbf{r}_{i+1}, L\mathbf{p}_i\rangle}{|L\mathbf{p}_i|^2}\,. \tag{61}$$

(2-4) and (60-61) together constitute the *conjugate residual* algorithm (of which GCR is a generalisation!) Since $\mathbf{p}_{i+1}$ depends only on $\mathbf{p}_i$ and $\mathbf{r}_{i+1}$, and $\mathbf{r}_{i+1}$ depends only on $\mathbf{p}_i$ and $\mathbf{r}_i$ the iterates $\mathbf{p}_{i+1}$ and $\mathbf{r}_{i+1}$ do not depend in any way on where in the loop they are calculated. Hence in the symmetric case the MO GCR(k) algorithm is equivalent to (2-4) and (60-61) for all i, and we have

Proposition 4

If LP^{-1} is symmetric then for any k, k′ the iterates $\mathbf{p}_i$, $\mathbf{x}_i$ and $\mathbf{r}_i$, i=1,2,... generated by MO GCR(k) are identical to those generated by MO GCR(k′), and in particular MO GCR(k) is equivalent to the full (ie unrestarted) GCR algorithm.

For our Helmholtz equation application in the Met Office P_1 is usually taken to be a block vertical preconditioner and P_i for i>1 some variety of ADI (but independent of i with P_i=P_2 for i≥2). The matrix LP_2^{-1} in both the PF and Adjoint Models is close to symmetric and in view of Proposition 4 we expect the convergence rate of GCR(k) for different values of k to be similar. Figure 1 shows the 2-norm (ie Euclidean norm) of the residual $|\mathbf{r}|$ against number of minor iterations[8] for GCR(1) (dotted), GCR(2) (dashed) and GCR(3), GCR(4) and GCR(5) (these together on the lowest line, indistinguishable to the naked eye) in the PF Model. Each of the five runs had the same right hand side $\mathbf{f}$ and same matrices L, $\{P_i\}$, differing only in k. See Appendix E for other details of code and settings used. Figure 2 repeats the experiment in the Adjoint.

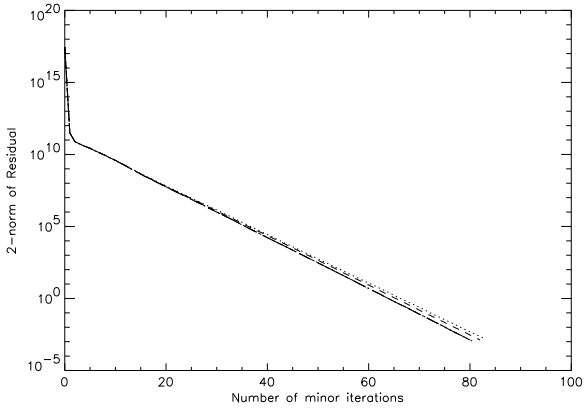

Figure 1: 2-norm of residual for GCR(1) (dotted), GCR(2) (dashed) MO GCR(3,4,5) (virtually coincident on the third line) against number of minor iterations in PF Model.

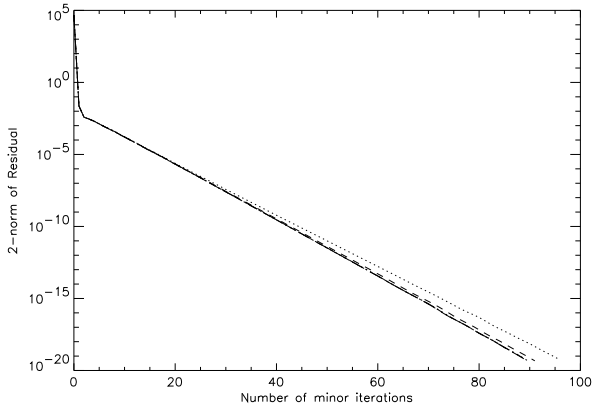

Figure 2: 2-norm of residual for GCR(1) (dotted), GCR(2) (dashed) MO GCR(3,4,5) (virtually coincident on the third line) against number of minor iterations in Adjoint Model.

Note that Proposition 4 is not true for EES's GCR(k) which restarts from scratch every k iterations. We can therefore expect MO GCR(k) to be more efficient than EES's GCR(k) if LP_2^{-1} is close to symmetric[9].

[8] A minor iteration being of the form $\{\mathbf{x}_i,\mathbf{r}_i,\mathbf{p}_i\}\rightarrow\ \{\mathbf{x}_{i+1},\mathbf{r}_{i+1},\mathbf{p}_{i+1}\}$, cf major iteration $\{\mathbf{x}_{mk},\mathbf{r}_{mk},\mathbf{p}_{mk}\}\rightarrow\ \{\mathbf{x}_{(m+1)k},\mathbf{r}_{(m+1)k},\mathbf{p}_{(m+1)k}\}$

[9] The number of major iterations required to reduce the 2-norm by 10^{12} in one run of the Adjoint Model for EES's GCR(k) and MO GCR(k), k=2,3 were as follows:

Timestep	EES GCR(2)	MO GCR(2)
500s	45	22
1000s	78	30
1500s	101	37
2000s	212	42

Timestep	EES GCR(3)	MO GCR(3)
500s	25	14
1000s	60	20
1500s	56	25
2000s	76	28

8.2 Rate of Convergence of MO GCR(k) for Symmetric LP^{-1}

If LP^{-1} is symmetric we can by virtue of Proposition 4 apply the well-known (see Section 6) characterisation of unrestarted GCR. If P is a fixed preconditioner repeated use of (3,16,60) gives

$$span\{L\mathbf{p}_0, .., L\mathbf{p}_j\} = span\{LP^{-1}\mathbf{r}_0, .., (LP^{-1})^i\mathbf{r}_0\} \tag{62}$$

Let $\mathcal{P}_i$ denote the set of real polynomials q_i of degree less than or equal to i such that $q_i(0)=1$. Inserting (62) into (38) with k=∞ shows that for each i

$$\mathbf{r}_i = q_i(LP^{-1})\mathbf{r}_0 \tag{63}$$

where $q_i \in \mathcal{P}_i$ is such that

$$|\mathbf{r}_i| = \min_{q_i \in \mathcal{P}_i} |q_i(LP^{-1})\mathbf{r}_0| \tag{64}$$

If LP^{-1} is symmetric

$$LP^{-1} = M\Lambda M^{-1} \tag{65}$$

where M is the orthogonal matrix of normalised eigenvectors of LP^{-1} and Λ is the diagonal matrix of eigenvalues of LP^{-1}. Hence

$$|\mathbf{r}_i| = \min_{q_i \in \mathcal{P}_i} |q_i(M\Lambda M^{-1})\mathbf{r}_0| \tag{66}$$

If we denote the eigenvectors of LP^{-1}

$$(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n) \equiv M$$

so that

$$M^T\mathbf{r}_0 = \begin{pmatrix} \langle \mathbf{v}_1, \mathbf{r}_0 \rangle \\ \langle \mathbf{v}_2, \mathbf{r}_0 \rangle \\ . \\ . \\ \langle \mathbf{v}_n, \mathbf{r}_0 \rangle \end{pmatrix}$$

and

$$q_i(\Lambda) = diag(q_i(\lambda_1), ..., q_i(\lambda_n))$$

where λ_1,..,λ_n are the eigenvalues of LP^{-1} corresponding to $\mathbf{v}_1$,...,$\mathbf{v}_n$ then

$$q_i(\Lambda)M^T\mathbf{r}_0 = \begin{pmatrix} q_i(\lambda_1)\langle \mathbf{v}_1, \mathbf{r}_0 \rangle \\ q_i(\lambda_2)\langle \mathbf{v}_2, \mathbf{r}_0 \rangle \\ . \\ . \\ q_i(\lambda_n)\langle \mathbf{v}_n, \mathbf{r}_0 \rangle \end{pmatrix} \tag{67}$$

The term on the right hand side of (66)

$$|q_i(M\Lambda M^{-1})\mathbf{r}_0| = |Mq_i(\Lambda)M^{-1}\mathbf{r}_0| = |q_i(\Lambda)M^T\mathbf{r}_0| \tag{68}$$

and so by (66,68,67)

$$|\mathbf{r}_i| = \min_{q_i \in \mathcal{P}_i} \left| \begin{pmatrix} q_i(\lambda_1)\langle \mathbf{v}_1, \mathbf{r}_0 \rangle \\ q_i(\lambda_2)\langle \mathbf{v}_2, \mathbf{r}_0 \rangle \\ . \\ . \\ q_i(\lambda_n)\langle \mathbf{v}_n, \mathbf{r}_0 \rangle \end{pmatrix} \right| \tag{69}$$

Recall that $\mathbf{r}_0 = \mathbf{f}$ where $\mathbf{f}$ is the right hand side of (1). From (69) it follows that if $\mathbf{f}$ only has a component parallel to i eigenvectors - ie if

$$\langle \mathbf{v}_j, \mathbf{f} \rangle = 0$$

for all but i values of j - then we could choose

$$q_i(\lambda) = \frac{\prod_{j=1}^{i}(\lambda_j - \lambda)}{\prod_{j=1}^{i}\lambda_j}$$

and so $|\mathbf{r}_i|$=0, ie MO GCR(k) would reduce the residual to zero in i iterations.

In the 'general' case where the right hand side $\mathbf{f}$ has a component parallel to more than a small subset of eigenvectors we may bound $|\mathbf{r}_i|$ in a similar way to [1] using Chebyshev polynomials. From (66)

$$|\mathbf{r}_i| = \min_{q_i \in \mathcal{P}_i} |q_i(M\Lambda M^{-1})\mathbf{r}_0| \leq \min_{q_i \in \mathcal{P}_i} \|Mq_i(\Lambda)M^{-1}\||\mathbf{r}_0| = \min_{q_i \in \mathcal{P}_i} \|q_i(\Lambda)\||\mathbf{r}_0| \tag{70}$$

and since

$$\|q_i(\Lambda)\| = \max_{j=1,\ldots,n} |q_i(\lambda_j)|$$

we have

$$|\mathbf{r}_i| \leq \min_{q_i \in \mathcal{P}_i} \max_{j=1,\ldots,n} |q_i(\lambda_j)||\mathbf{r}_0| \tag{71}$$

If the eigenvalues λ_1,..,λ_n of LP^{-1} lie on one side of the origin (ie LP^{-1} is definite) and are 'densely' distributed in the range $[\lambda_{min},\lambda_{max}]$ then

$$\min_{q_i \in \mathcal{P}_i} \max_{j=1,\ldots,n} |q_i(\lambda_j)| \simeq \min_{q_i \in \mathcal{P}_i} \max_{\lambda \in [\lambda_{min},\lambda_{max}]} |q_i(\lambda)| \tag{72}$$

The monic ith order polynomial q_i which is the solution to the right hand term of (72) is known (see Theorem B.1 of [1]) to be

$$P_i[\lambda_{min}, \lambda_{max}](\lambda) \equiv \frac{T_i\left(\frac{\lambda_{min}+\lambda_{max}-2\lambda}{\lambda_{max}-\lambda_{min}}\right)}{T_i\left(\frac{\lambda_{min}+\lambda_{max}}{\lambda_{max}-\lambda_{min}}\right)} \tag{73}$$

where

$$T_i(x) = \frac{1}{2}[(x+\sqrt{x^2-1})^i + (x-\sqrt{x^2-1})^i] \tag{74}$$

is the ith Chebyshev polynomial. Furthermore

$$max_{\lambda \in [\lambda_{min},\lambda_{max}]}|P_i[\lambda_{min}, \lambda_{max}](\lambda)| = \frac{1}{T_i\left(\frac{\lambda_{min}+\lambda_{max}}{\lambda_{max}-\lambda_{min}}\right)} = 2\frac{\sigma^i}{1+\sigma^{2i}} \tag{75}$$

where

$$\sigma = \frac{1-\sqrt{\lambda_{min}/\lambda_{max}}}{1+\sqrt{\lambda_{min}/\lambda_{max}}} \tag{76}$$

Combining (71, 72, 75) we have

$$\frac{|\mathbf{r}_i|}{|\mathbf{r}_0|} \leq 2\frac{\sigma^i}{1+\sigma^{2i}} \tag{77}$$

and hence the average fraction by which $|\mathbf{r}_i|$ is reduced each iteration, is

$$\left(\frac{|\mathbf{r}_i|}{|\mathbf{r}_0|}\right)^{\frac{1}{i}} \leq 2^{1/i}\frac{\sigma}{(1+\sigma^{2i})^{1/i}} \tag{78}$$

which $\rightarrow \sigma$ as i $\rightarrow \infty$

These results are tested against real data in Section 11 below[10].

[10]Note that the results (73-78) are only valid for positive definite LP^{-1}. Extension to the indefinite case is possible but not pursued here.

8.3 Preconditioning GCR(k) to Accelerate Convergence

It follows from (77) etc that to obtain fastest convergence we need to choose P_i so that the eigenvalues of LP_i^{-1} are real and positive and such that

$$\frac{\lambda_{min}(LP_i^{-1})}{\lambda_{max}(LP_i^{-1})} \tag{79}$$

is as large (ie as close to 1) as possible. That is, we seek P_i such that $\mathrm{LP}_i^{-1} \approx I$. Currently for the Helmholtz-type equation of the full model and 4D-Var the Met Office uses for P_1 a block vertical preconditioner and for P_i with i>1 a variety of ADI.

9 Cost of the MO GCR(k)

The MO GCR(k) stated in its simplest form in Section 2, with preconditioning in Section 5 and in the form used in the code in Appendix D consists of a double loop in which each major iteration comprises the k minor iterations shown in the left hand column of Table 1. Let the dimension of L be N. In the first right hand column is listed which of the four types of significant operation is involved - scalar products ('sp', 2N flops), vector plus scalar times vector ('vpstv', 2N flops), preconditioning or matrix vector products L(v), ie calls to GCR_elliptic_operator - and in the last column for the first two types their cost in floating point operations.

Do while $i = 0, ..., k-1$ *and convergence not obtained*

	Operation(s)	*flops*
Compute $\langle L(\mathbf{p}_i), L(\mathbf{p}_i)\rangle$	*sp*	$2N$
$\beta = -\dfrac{\langle \mathbf{r}_i, L(\mathbf{p}_i)\rangle}{\langle L(\mathbf{p}_i), L(\mathbf{p}_i)\rangle}$	*sp*	$2N$
$\mathbf{x}_{i+1} = \mathbf{x}_i + \beta \mathbf{p}_i$	*vpstv*	$2N$
$\mathbf{r}_{i+1} = \mathbf{r}_i + \beta L(\mathbf{p}_i)$	*vpstv*	$2N$
$\mathbf{q} = P^{-1}(\mathbf{r}_{i+1})$	*precond*	C'
$CalculateL(\mathbf{q})$	*mv product*	C
$\forall l = 0, i \;\; \alpha^l = -\dfrac{\langle L(\mathbf{q}), L(\mathbf{p}_l)\rangle}{\langle L(\mathbf{p}_l), L(\mathbf{p}_l)\rangle}$	$i+1$ *sp's*	$2(i+1)N$
$\mathbf{p}_{i+1} = \mathbf{q} + \sum_{l=0}^{i} \alpha^l \mathbf{p}_l$	$i+1$ *vpstv's*	$2(i+1)N$
$L(\mathbf{p}_{i+1}) = L(\mathbf{q}) + \sum_{l=0}^{i} \alpha^l L(\mathbf{p}_l)$	$i+1$ *vpstv's*	$2(i+1)N$

end do while

Table 1: **Operation Count for MO GCR(k)**

Thus if the cost of calling GCR_elliptic_operator is C flops and the cost of the precondioner C′ flops the cost of each major iteration of GCR(k) is

$$k \times (8N + C + C') + \sum_{i=0}^{k-1} 6(i+1)N = k[(11+3k)N + C + C'] \; flops \tag{80}$$

Plainly (in view of the $3Nk^2$ term) this is always more than k times the cost of GCR(1), though approaches being k times the cost of GCR(1) if the dominant cost in GCR(k) is that of calling the preconditioner plus GCR_elliptic_operator.

Experiments have been run to assess how superlinear in k the cost of GCR(k) is in the Met Office Helmholtz solver applications. Unpreconditioned GCR(k) was run for various values of k, and the CPU time per major iteration calculated. The figures in Table 2 are for four processors (2×2). The maximum memory used was 36.05 MWords, 36.18 MWords, 36.52 MWords, 37.0586 MWords, 37.5898 MWords for k=1,2,3,4,5 respectively. At the resolution used $N \approx 2.5 \times 10^5$.

k	*CPU Time per* 100 *major iterations*	*Mean CPU Time per* 100 *iterations*
1	$18.36s$	$18.36s$
2	$37.44s$	$18.72s$
3	$56.26s$	$18.75s$
4	$77.22s$	$19.31s$
5	$97.95s$	$19.59s$

Table 2: **Timings of Unpreconditioned GCR(k)**

From these figures it is clear that the dominant cost in the unpreconditioned case is the call to GCR_elliptic_operator and is only barely superlinear in k. In fact, they suggest C∼200N.

Note that similar formulae to (80) can be written down for EES GCR(k), Orthomin(k) etc (see [2]), differing only in the term premultiplying the N, ie with the same dominant cost per iteration. The main issue in comparing these methods is therefore the effectiveness of each iteration in reducing the magnitude of the residual. In this regard, from Section 8.1 we expect in the nearly symmetric case little difference between MO GCR(k) and Orthomin(k), since for all k they both approximate GCR(∞). EES GCR(k) does not have this property and in the nearly symmetric case is much less efficient, see footnote (9) on page 20.

10 The Eigenvalues of L in the PF and Adjoint Models

In view of the results of Section 8 and for other reasons we are interested in the extremal eigenvalues of the matrices L and LP^{-1} as they occur in the Helmholtz-type equations in Met Office models. The extremal eigenvalues of L were calculated for the PF and Adjoint Models for timesteps of 50s to 3600s at intervals of 50s using the methods described in Appendices C.1 and C.3. Since in both the PF and Adjoint Model L is close to symmetric we expect most eigenvalues to be real, and this was found to be the case (all the most extremal ones were).
The eigenvalues of L of smallest modulus are well behaved as a function of timestep Δt; they are always negative in the PF model and positive in the Adjoint, see Figures 3 and 4.
The largest positive and largest negative eigenvalues are more interesting. Figures 6, 7, 8 and 9 show them in the PF and Adjoint Models. The number of positive eigenvalues in the PF model is shown for timesteps up to 1800s in Figure 5; at least for timesteps in this range exactly the same number of negative eigenvalues were found for corresponding timesteps in the Adjoint.
The most striking feature of Figures 6, 7, 8 and 9 are the spikes at certain timesteps; increasing the resolution around them (ie decreasing the interval between timesteps from 50s to 1s to 0.01s etc) shows that the spikes are in fact infinite. We know from Gersgorin's theorem (see eg [5]) that if all the elements of L(Δt) are uniformly bounded for timesteps in some interval [a,b] that the eigenvalues

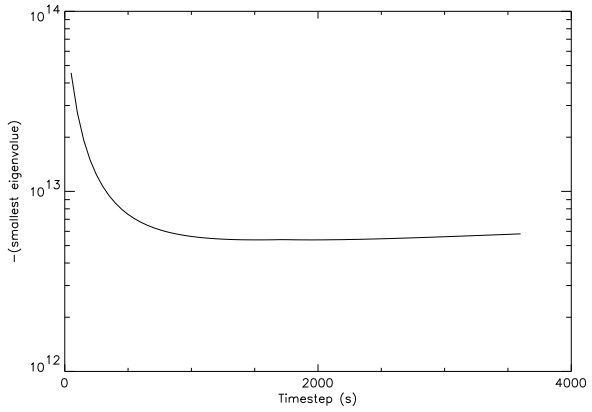

Figure 3: Modulus of the smallest eigenvalue of L in the PF model.

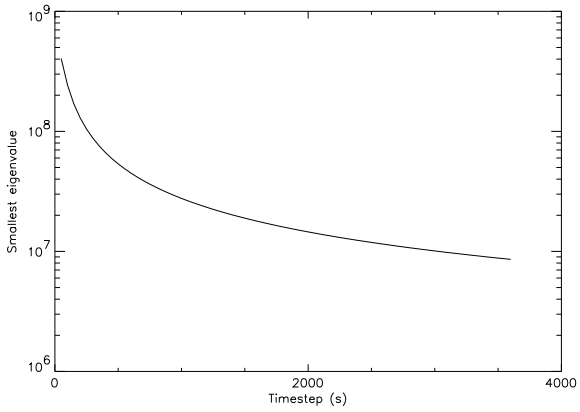

Figure 4: Smallest eigenvalue of L in the Adjoint model.

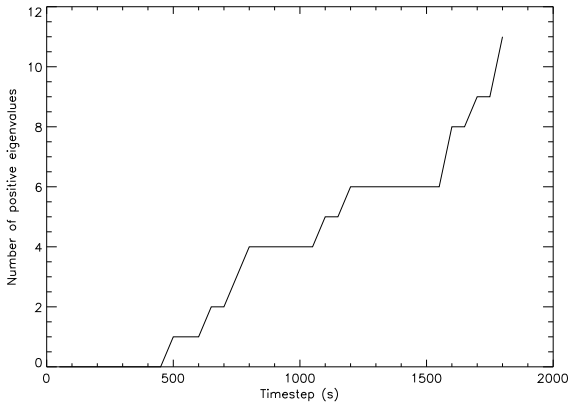

Figure 5: Number of positive eigenvalues of L in PF model for timesteps up to 1800s.

of L(Δt) must also be uniformly bounded for all timesteps in [a,b], and hence some element(s) of L(Δt) must be going infinite at those timesteps Δt where the spikes occur.

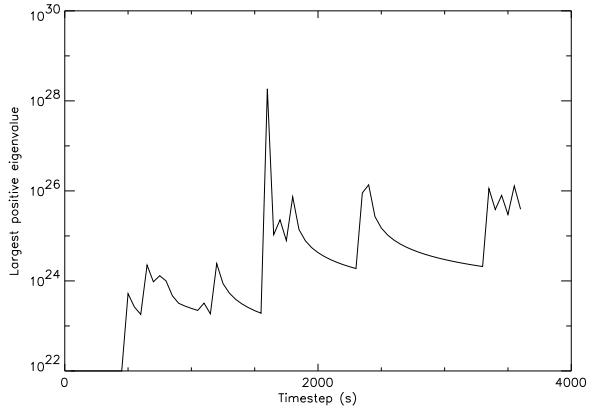

Figure 6: Largest positive eigenvalue of L in the PF model. For timesteps less than 450s all eigenvalues are negative.

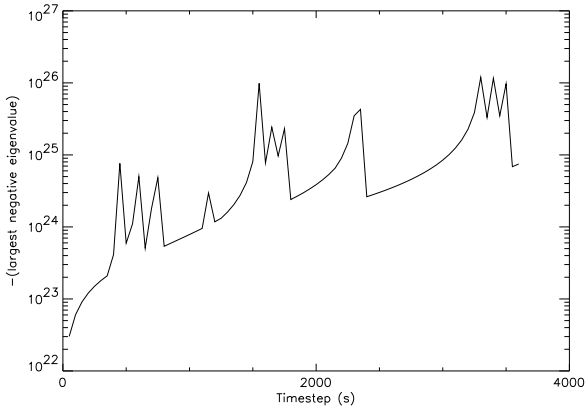

Figure 7: Modulus of the largest negative eigenvalue of L in the PF model.

The coefficients of L(Δt) as given in an appendix to [13] show that the only term which can cause such singularities is[11]

$$G = 1 - \alpha_2 \alpha_3 c_p (\Delta t)^2 \delta_{2r} \theta \delta_r \Pi \tag{81}$$

[11] Later editions of the code replaced the α_3 by α_4, but the version given in (81) was used for all the runs in this report.

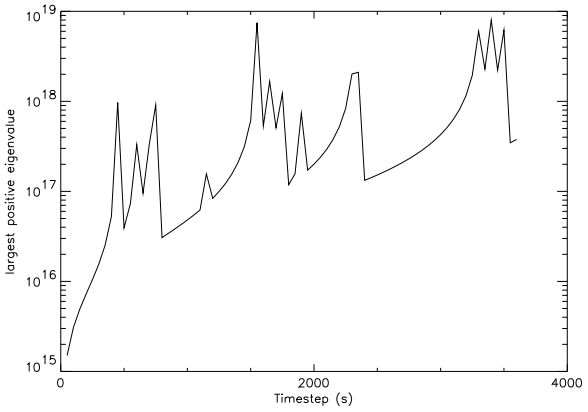

Figure 8: Largest positive eigenvalue of L in the Adjoint model.

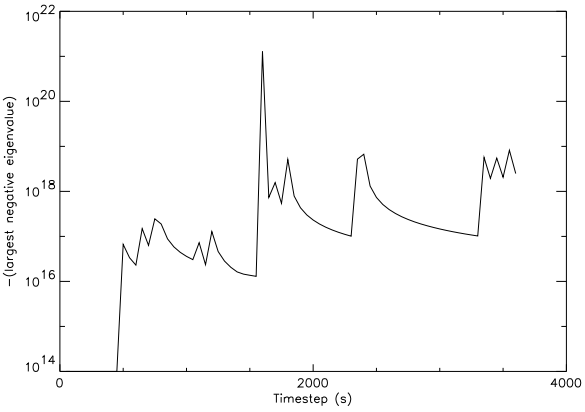

Figure 9: Modulus of the largest negative eigenvalue of L in the Adjoint model. For timesteps less than 450s all eigenvalues are positive.

which appears in the denominators of C_z and C_{zz} in the PF Model and of C_{zz} in the Adjoint Model. For the linearisation state (without physics) with which we were working (81) implied that G would vanish at 22 quadruples (Δt,i,j,k) of which the 17 with timesteps less than 3600s are listed in Table 3[12].

$Timestep\,(s)$	(i,j,k)
453.04945	$(90,64,1)$
615.09101	$(12,54,1)$
720.57438	$(81,7,1)$
758.35657	$(70,7,1)$
1089.7348	$(76,6,4)$
1172.3826	$(80,28,2)$
1554.5488	$(63,57,1)$
1599.9716	$(80,19,1)$
1674.0431	$(11,55,1)$
1786.7948	$(4,28,1)$
1790.986	$(34,55,1)$
1913.2429	$(76,5,2)$
2335.9212	$(70,50,1)$
2387.9493	$(87,35,1)$
3324.1424	$(9,31,1)$
3420.254	$(5,55,1)$
3528.4311	$(80,21,1)$

Table 3: **Timesteps (less than 3600s) and corresponding gridpoints where G vanishes**

These timesteps correspond to the spikes in Figures 6, 7, 8 and 9 and in fact comparing Table 3 with Figure 5 we see that (at least for timesteps up to 1800s) whenever the timestep Δt passes through a value where G vanishes at some gridpoint the number of positive eigenvalues in the PF Model (or negative eigenvalues in the Adjoint) increases by 1.

[12]The situation is dramatically different if the linearisation state includes physics. In that case G vanishes for some timestep at around half of all points in the first couple of layers, so even at climate resolution a table corresponding to Table 3 would be many thousands of lines long.

There is an interesting link between G vanishing, definiteness of L and ellipticity of the original Helmholtz problem

$$\mathcal{L}(\psi) = f \ subject\ to\ bc's\ on\ \psi\ at\ top\ and\ bottom \tag{82}$$

of which (1) is a discretisation. If

$$\begin{aligned}\mathcal{L} = C_{\lambda\lambda}\delta_{\lambda\lambda} + C_{\lambda\phi}\delta_{\lambda\phi} + C_{\lambda\nu}\delta_{\lambda\nu} + \\ C_{\phi\lambda}\delta_{\phi\lambda} + C_{\phi\phi}\delta_{\phi\phi} + C_{\phi\nu}\delta_{\phi\nu} + \\ C_{\nu\lambda}\delta_{\nu\lambda} + C_{\nu\phi}\delta_{\nu\phi} + C_{\nu\nu}\delta_{\nu\nu} + \ lower\ order\ terms\end{aligned}$$

then by definition $\mathcal{L}$ is elliptic if the matrix

$$\begin{pmatrix} C_{\lambda\lambda} & C_{\lambda\phi} & C_{\lambda\nu} \\ C_{\phi\lambda} & C_{\phi\phi} & C_{\phi\nu} \\ C_{\nu\lambda} & C_{\nu\phi} & C_{\nu\nu} \end{pmatrix} \tag{83}$$

is everywhere definite. In Section 15 of [12] it is shown that if orographic terms are neglected and other mild approximations made that this is so if and only if G is everywhere positive. We may relate definiteness of (83) to definiteness of L using ideas similar to those of Section 20.5 of [4]. If $\mathcal{L}$ is self-adjoint (and it nearly is) then solving (82) is equivalent to finding stationary values over ψ of

$$\begin{aligned}I(\psi) = \frac{1}{2}\int (C_{\lambda\lambda}(\delta_\lambda\psi)^2 + C_{\phi\phi}(\delta_\phi\psi)^2 + C_{\nu\nu}(\delta_\nu\nu)^2 + \\ (C_{\lambda\phi} + C_{\phi\lambda})\delta_\lambda\psi\delta_\phi\psi + (C_{\lambda\nu} + C_{\nu\lambda})\delta_\lambda\psi\delta_\nu\psi + (C_{\phi\nu} + C_{\nu\phi})\delta_\phi\psi\delta_\nu\psi)dv\end{aligned} \tag{84}$$

If $\mathcal{L}$ is elliptic the stationary value is a minimum. We may discretise (84) in such a way that it transforms into: find vector $\mathbf{s}$ which minimises

$$I_\Delta(\mathbf{s}) = \sum f_i s_i - \frac{1}{2}\sum L_{ij}s_i s_j \tag{85}$$

where L is the matrix in (1). But it is clear that in this formulation (which is only valid if $\mathcal{L}$ is elliptic, which as we have seen requires G positive everywhere) L must be negative definite. We reran the jobs used to create Figures 6, 7 and 3, with the one difference that $G_{ijk}(\Delta t)$ was replaced by $\max(0.9, G_{ijk}(\Delta t))$. The extremal eigenvalues were then recalculated. The smallest one were scarcely affected, but there were now no positive eigenvalues, and the largest negative one (Figure 10) is now a smooth function of Δt (cf Figure 7).

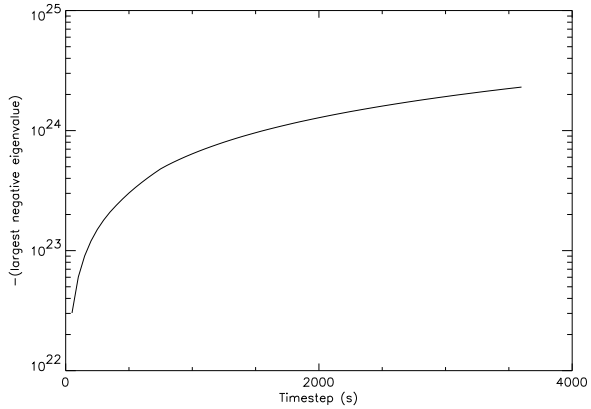

Figure 10: Modulus of largest eigenvalue of L in the PF model with G limited below to 0.9 (as discussed in Section 11 this forces all eigenvalues to be negative). cf Figure 7.

11 The Eigenvalues of LP^{-1} and the rate of convergence of MO GCR(k) in the PF and Adjoint Models

Using the methods described in Appendices C.2 and C.4 the largest and smallest eigenvalues of the preconditioned matrices LP^{-1} in the PF and Adjoint models, with P^{-1} the XZ ADI preconditioner, were calculated for timesteps of 50s to 3600s at intervals of 50s. Plainly if the preconditioning was perfect all eigenvalues would be 1.0. In fact all the eigenvalues have positive real part and almost all are real (those of minimum and maximum modulus always are). The largest eigenvalue of the preconditioned matrix (exhibited for the Adjoint in Figure 11) is always close to 1.0.

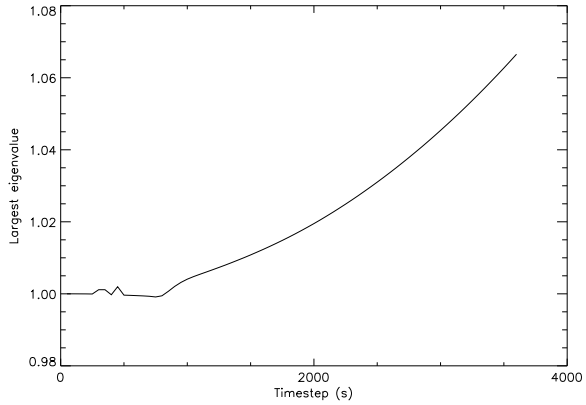

Figure 11: Largest eigenvalue of LP^{-1} (P=XZ ADI) in the Adjoint model.

Figures 12 and 13 show the smallest eigenvalues in the PF and Adjoint Models respectively. In the PF model the pseudo-timestep was fixed at 0.006s for all real timesteps while in the Adjoint the pseudo-timestep was calculated according to

$$\text{ADI pseudo} - \text{timestep} = -150000s \times \frac{\text{real timestep}}{1800}$$

which is why the Adjoint is more sucessful at keeping the smallest eigenvalue close to 1.0 at smaller real timesteps.

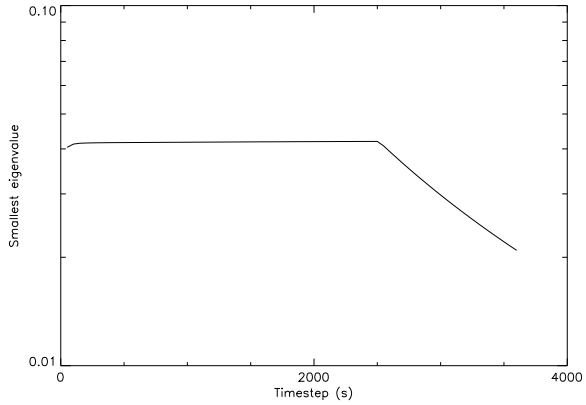

Figure 12: Smallest eigenvalue of LP^{-1} (P=XZ ADI) in the PF model.

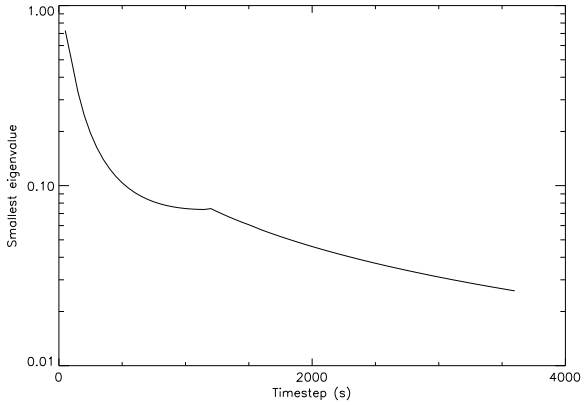

Figure 13: Smallest eigenvalue of LP^{-1} (P=XZ ADI) in the Adjoint model.

We may use Figures 11 and 13 and equation (78) to predict the rate of convergence of MO GCR(k). Runs similar to those which generated the dashed (ie GCR(2)) curve in Figure 2 were made with real timesteps of 50,100,200,400,800,1600 and 3200s. In each case (as in Figure 2) the curves were fairly straight after the initial block vertical iteration (XZ ADI preconditioning was used for second

and subsequent minor iterations), ie

$$\frac{log_{10}|\mathbf{r}_{i+m}| - log_{10}|\mathbf{r}_m|}{i}$$

is fairly independent of i and m so long as m is greater than about 2. It follows that

$$\left(\frac{|\mathbf{r}_{i+m}|}{|\mathbf{r}_m|}\right)^{1/i} \tag{86}$$

is fairly independent of i and m so long as m is greater than about 2. Values of the quantity (86) for the seven timesteps listed above were found to be as given in table 4.

$Timestep\ (s)$	$\left(\frac{\lvert\mathbf{r}_{i+m}\rvert}{\lvert\mathbf{r}_m\rvert}\right)^{1/i}$	$\pm$
50	0.07	0.01
100	0.15	0.015
200	0.3	0.05
400	0.42	0.08
800	0.535	0.005
1600	0.635	0.05
3200	0.73	0.02

Table 4:

If equality is achieved in (78) then for large enough i

$$\left(\frac{|\mathbf{r}_{i+m}|}{|\mathbf{r}_m|}\right)^{1/i} \simeq \frac{1-\sqrt{\lambda_{min}/\lambda_{max}}}{1+\sqrt{\lambda_{min}/\lambda_{max}}} \tag{87}$$

Using the data in Figures 11 and 13 we may readily calculate the quantity on the right hand side of (87) as given in table 5.

$Timestep\ (s)$	$\frac{1-\sqrt{\lambda_{min}/\lambda_{max}}}{1+\sqrt{\lambda_{min}/\lambda_{max}}}$
50	0.079461
100	0.17387
200	0.336746
400	0.47867
800	0.560668
1600	0.616982
3200	0.714398

Table 5:

The reduction factor $\left(\frac{|\mathbf{r}_{i+m}|}{|\mathbf{r}_m|}\right)^{1/i}$ as measured form real data (Table 4) is plotted as the dotted curve and our prediction $\frac{1-\sqrt{\lambda_{min}/\lambda_{max}}}{1+\sqrt{\lambda_{min}/\lambda_{max}}}$ of it according to (87) (Table 5) as the dashed curve in Figure 14.

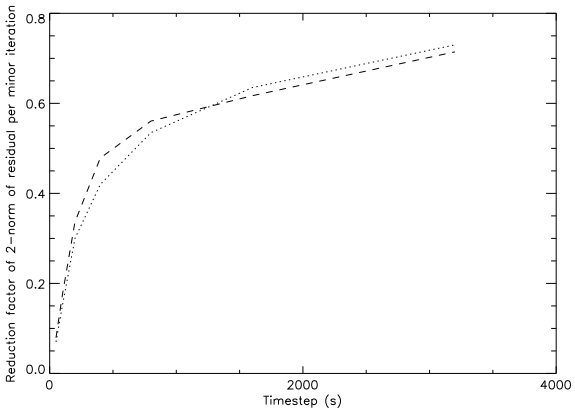

Figure 14: Predicted (dotted) and actual (dashed) reduction in 2-norm of residual in the Adjoint Model, various timesteps.

Figures 15 - 18 show the number of major GCR(2) iterations for the PF and Adjoint Models to converge according to the criterion that the infinity norm of the residual is reduced to respectively 10^{-12} and 10^{-24}. In Figures 15 and 17 the block vertical preconditioner is used throughout and in Figures 16 and 18 block vertical preconditioning is used on the first minor iteration and XZ ADI preconditioning subsequently. See above remark concerning the different way in which the ADI pseudo-timestep was chosen in The PF and Adjoint Models and its effect on convergence for small real timesteps.

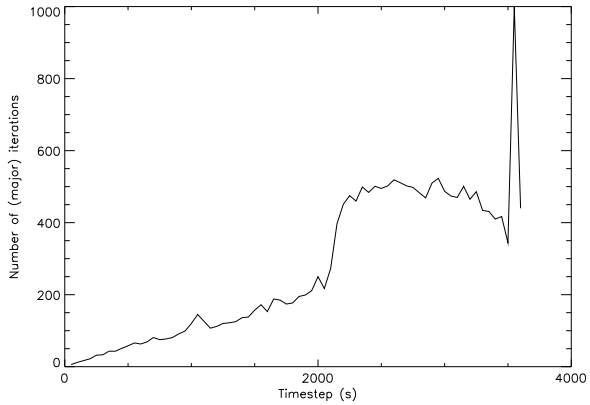

Figure 15: Number of GCR(2) iterations required to reduce the infinity norm of the residual to 10^{-12} in PF model using block vertical preconditioning.

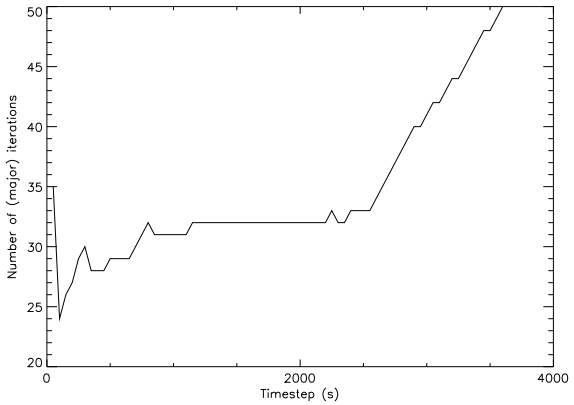

Figure 16: Number of GCR(2) iterations required to reduce the infinity norm of the residual to 10^{-12} in PF model using XZ ADI preconditioning.

Given our success, as illustrated in Figure 14, in predicting convergence rates from knowledge of the extremal eigenvalues we can attempt to predict the number of iterations to reach a given convergence level (ie Figures 15 - 18) also from knowledge of the extremal eigenvalues. In applying this to Figure 18 for example there are though now two complications: firstly that the eigenvalues in Figures 13 and 11 are for XZ ADI while the first minor iteration in each run of Figure 18 has block vertical preconditioning, and secondly that the criterion is to stop iterating when the infinity norm of the residual descends to 10^{-24} while our formulae (77) and (78) are for a relative reduction in 2-norm. We shall therefore need to approximate: typically the 2-norm reduction required from the second iteration to reduce the infinity norm of the residual in the Adjoint to 10^{-24} is a factor of about $\delta \equiv 10^{-12}$. If equality is achieved in (77) this means that the number of minor iterations required to converge is i+1 (the plus one because of the initial block vertical minor iteration) where i satisfies

$$\frac{2\sigma^i}{1+\sigma^{2i}} \simeq \delta \tag{88}$$

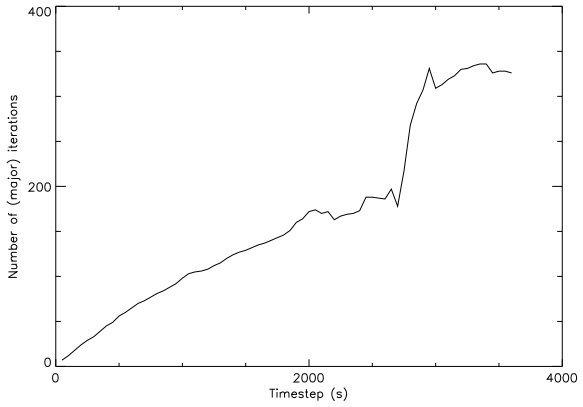

Figure 17: Number of GCR(2) iterations required to reduce the infinity norm of the residual to 10^{-24} in Adjoint model using block vertical preconditioning.

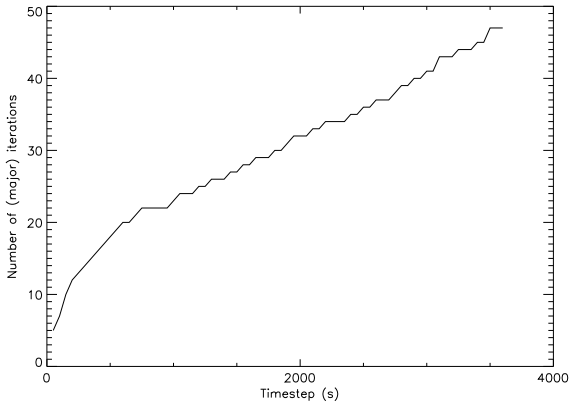

Figure 18: Number of GCR(2) iterations required to reduce the infinity norm of the residual to 10^{-24} in Adjoint model using XZ ADI preconditioning.

where σ is as defined in (76). Approximating the left hand side of (88) for $\sigma \ll 1$ by $2\sigma^i$ and neglecting $\log_{10}2$ compared with 12 we expect convergence in approximately

$$\frac{1}{2}\left(\frac{-12}{log_{10}\sigma}+1\right) \tag{89}$$

major GCR(2) iterations. Using the data from Figures 13 and 11 this quantity is plotted in Figure 19. In fact, the 2-norm reduction required from the second iteration to reduce the infinity norm of the residual in the Adjoint to 10^{-24} is rather less than 10^{-12} for shorter timesteps and more for longer ones, and had this been taken into account the agreement between Figures 18 and 19 would have been even better.

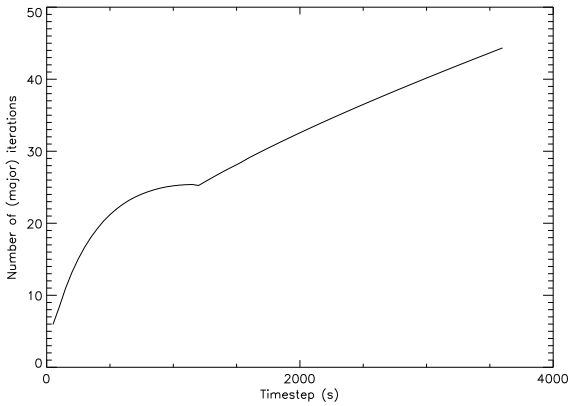

Figure 19: Number of GCR(2) iterations to converge using ADI preconditioning in Adjoint model as predicted by (89), cf Figure 18.

12 Summary

This report has derived the main properties of the variety of GCR(k) used in the Met Office and compared it with similar algorithms.

MO GCR(k) is only a "standard" algorithm when k=1, when it coincides with Orthomin(1), but it was shown in §3 how for any k it can be placed within the unifying framework of Restarted/Truncated Petrov-Galerkin-Krylov Algorithms.

Orthomin(1) can be obtained by discretising a 2nd order damped wave equation; it was shown in §4 that the generalisation to a (k+1)th order damped wave equation yields Orthomin(k) rather than MO GCR(k).

MO GCR(k) shares some of the properties of [4]'s ('EES') GCR(k), most notably that positive definiteness of the symmetric part of LP^{-1}, where P is the preconditioner, guarantees convergence for any $k \geq 1$ (§6).

Clearly the overall efficiency of an iterative method depends both on the effectiveness of each iteration in reducing the error and on the cost per iteration. In the first regard, a key difference with EES GCR(k) is that MO GCR(k) is independent of k for LP^{-1} symmetric (§8.1). This is an advantage for applications where LP^{-1} is close to symmetric since then convergence is nearly as fast as with GCR(∞). In the second regard, the cost per iteration is derived in §9 and shown in Met Office applications to be dominated by the application of L and the preconditioner, and so is similar for EES GCR(k), Orthomin(k) and MO GCR(k).

We have also examined the eigenvalues of L and LP^{-1}, where L is the Helmholtz operator and P a preconditioner. It was verified (§10) that L is definite if and only if $\mathcal{L}$ is elliptic and if and only if G as defined in eqn (81) nowhere vanishes. It has also been shown (§11) that in Met Office applications an estimate of convergence rate (derived in §8.2) based on measured eigenvalues of LP^{-1} and assuming LP^{-1} symmetric predicts actual convergence rates with good accuracy.

Appendices

A Discretising The Damped Wave Equation (15) with k=3 yields Orthomin(2)

Discretising (15) with k=3 (centering first and third derivatives on n-$\frac{1}{2}$ and the second derivative on n) gives

$$\frac{\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}_{n-2} - 3\boldsymbol{\phi}_n + 3\boldsymbol{\phi}_{n-1}}{(\Delta\tau)^3} + \frac{1}{T_2(n)}\frac{\boldsymbol{\phi}_{n+1} + \boldsymbol{\phi}_{n-1} - 2\boldsymbol{\phi}_n}{(\Delta\tau)^2} + \frac{1}{T_1(n)}\frac{\boldsymbol{\phi}_n - \boldsymbol{\phi}_{n-1}}{\Delta\tau} = L(\boldsymbol{\phi}_n) - \mathbf{f} \equiv \mathbf{r}_n \tag{90}$$

so

$$\boldsymbol{\phi}_{n+1} = \beta_n \mathbf{r}_n + a_n \boldsymbol{\phi}_n + b_n \boldsymbol{\phi}_{n-1} + (1 - a_n - b_n)\boldsymbol{\phi}_{n-2} \tag{91}$$

where

$$\beta_n = \frac{(\Delta\tau)^3}{1 + \frac{\Delta\tau}{T_2(n)}}$$

$$a_n = \frac{\beta_n}{(\Delta\tau)^3}\left(3 + \frac{2\Delta\tau}{T_2(n)} - \frac{(\Delta\tau)^2}{T_1(n)}\right)$$

$$b_n = \frac{\beta_n}{(\Delta\tau)^3}\left(-3 - \frac{\Delta\tau}{T_2(n)} + \frac{(\Delta\tau)^2}{T_1(n)}\right)$$

Our objective is to choose β_n, a_n and b_n to minimise the 2-norm of the residual error $\mathbf{r}_{n+1}$.

For $\beta_n \neq 0$ define

$$\mathbf{p}_n = \frac{\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}_n}{\beta_n}$$

so that

$$\boldsymbol{\phi}_{n+1} = \beta_n \mathbf{p}_n + \boldsymbol{\phi}_n \tag{92}$$

and (91) is equivalent to

$$\mathbf{p}_n = \mathbf{r}_n + \frac{\beta_{n-1}}{\beta_n}(a_n - 1)\mathbf{p}_{n-1} + \frac{\beta_{n-2}}{\beta_n}(a_n + b_n - 1)\mathbf{p}_{n-2}$$

It is convenient to define

$$A_n = \frac{\beta_{n-1}}{\beta_n}(a_n - 1)$$

$$B_n = \frac{\beta_{n-2}}{\beta_n}(a_n + b_n - 1)$$

so that

$$\mathbf{p}_n = \mathbf{r}_n + A_n \mathbf{p}_{n-1} + B_n \mathbf{p}_{n-2} \tag{93}$$

Taking L of (92) gives

$$\mathbf{r}_{n+1} = \beta_n L(\mathbf{p}_n) + \mathbf{r}_n \tag{94}$$

and our objective is to find β_n, A_n and B_n to minimise the 2-norm of the residual error $\mathbf{r}_{n+1}$ in (94). We do this by solving

$$\frac{\partial}{\partial \beta_n}\langle \mathbf{r}_{n+1}, \mathbf{r}_{n+1}\rangle = 0 \tag{95}$$

$$\frac{\partial}{\partial A_{n+1}}\langle \mathbf{r}_{n+2}, \mathbf{r}_{n+2}\rangle = 0 \tag{96}$$

$$\frac{\partial}{\partial B_{n+1}}\langle \mathbf{r}_{n+2}, \mathbf{r}_{n+2}\rangle = 0 \tag{97}$$

Using (93) and (94) it follows that (95) implies

$$\langle \mathbf{r}_{n+1}, L(\mathbf{p}_n)\rangle = 0 \tag{98}$$

that (96) implies

$$\langle \mathbf{r}_{n+2}, L(\mathbf{p}_n)\rangle = 0 \tag{99}$$

and that (97) implies

$$\langle \mathbf{r}_{n+2}, L(\mathbf{p}_{n-1})\rangle = 0 \tag{100}$$

Substituting (93) and (94) into (98-100) it follows that

$$\langle L(\mathbf{p}_{n+1}), L(\mathbf{p}_n)\rangle = \langle L(\mathbf{p}_{n+1}), L(\mathbf{p}_{n-1})\rangle = 0$$

and after some manipulation that

$$\beta_n = -\frac{\langle \mathbf{r}_n, L(\mathbf{p}_n)\rangle}{|L(\mathbf{p}_n)^2|} \tag{101}$$

$$A_{n+1} = -\frac{\langle L(\mathbf{r}_{n+1}), L(\mathbf{p}_n)\rangle}{|L(\mathbf{p}_n)^2|} \tag{102}$$

$$B_{n+1} = -\frac{\langle L(\mathbf{r}_{n+1}), L(\mathbf{p}_{n-1})\rangle}{|L(\mathbf{p}_{n-1})^2|} \tag{103}$$

This gives the algorithm:

Given $\mathbf{r}_n$, $\mathbf{p}_n$, $\mathbf{p}_{n-1}$

Calculate β_n by (101)
Calculate $\mathbf{r}_{n+1}$ by (94)
Calculate A_{n+1} and B_{n+1} by (102,103)
Calculate $\mathbf{p}_{n+1}$ by (93)

which is Orthomin(2).

B Proofs of Results in Section 5

Proof of Relations (19)-(25)

(19) follows from (17) and the definition of b^i_j in (18)
(20) is vacuously true if i=mk. Suppose true for all mk $\leq$ j < i $\leq$ t, some $mk \leq t < (m+1)k$. From (3)

$$\langle \mathbf{r}_{t+1}, L\mathbf{p}_j \rangle = \langle \mathbf{r}_t, L\mathbf{p}_j \rangle - a_t \langle L\mathbf{p}_t, L\mathbf{p}_j \rangle$$

if j<t the right hand side is zero by hypothesis and (19), if j=t the right hand side is zero by definition (4) of a_t, and the result follows by induction.
(21): By (17)

$$\langle \mathbf{r}_i, L\mathbf{p}_i \rangle = \langle \mathbf{r}_i, LP_i^{-1}\mathbf{r}_i \rangle + b^{i-1}_{mk}\langle \mathbf{r}_i, L\mathbf{p}_{mk} \rangle + ... + b^{i-1}_{i-1}\langle \mathbf{r}_i, L\mathbf{p}_{i-1} \rangle$$

and all terms on the the right hand side except $\langle \mathbf{r}_i, LP_i^{-1}\mathbf{r}_i \rangle$ vanish by (20).
(22): From (17) we have

$$P_j^{-1}\mathbf{r}_j = \mathbf{p}_j - b^{j-1}_{mk}\mathbf{p}_{mk} - ... - b^{j-1}_{j-1}\mathbf{p}_{j-1}$$

Premultiplying this expression by L, taking the inner product with $\mathbf{r}_i$ for i > j and using (20) gives (22).
(23) is automatically true when i=mk. Suppose true for all mk$\leq$j$\leq$i $\leq$ t<(m+1)k. We need to show

$$\langle \mathbf{r}_j, L\mathbf{p}_{t+1} \rangle = \langle \mathbf{r}_{mk}, L\mathbf{p}_{t+1} \rangle \text{ any } mk \leq j \leq i \leq t+1$$

which again is immediately true if j=mk, so suppose true for all j $\leq$k. By (3), (19) and hypothesis

$$\langle \mathbf{r}_{k+1}, L\mathbf{p}_{t+1} \rangle = \langle \mathbf{r}_k - a_k L\mathbf{p}_k, L\mathbf{p}_{t+1} \rangle = \langle \mathbf{r}_k, L\mathbf{p}_{t+1} \rangle = \langle \mathbf{r}_{mk}, L\mathbf{p}_{t+1} \rangle$$

(24) follows by taking the scalar product of L$\times$(17) with L$\mathbf{p}_{i+1}$ and using (19)
(25): Using (17) and (19)

$$|L\mathbf{p}_i|^2 = |LP_i^{-1}\mathbf{r}_i|^2 + (b^{i-1}_{mk}L\mathbf{p}_{mk})^2 + ... + (b^{i-1}_{i-1}L\mathbf{p}_{i-1})^2 + 2\langle LP_i^{-1}\mathbf{r}_i, b^{i-1}_{mk}L\mathbf{p}_{mk} + ... + b^{i-1}_{i-1}L\mathbf{p}_{i-1} \rangle$$

$$= |LP_i^{-1}\mathbf{r}_i|^2 - \frac{\langle LP_i^{-1}\mathbf{r}_i, L\mathbf{p}_{mk}\rangle^2}{|L\mathbf{p}_{mk}|^2} - ... - \frac{\langle LP_i^{-1}\mathbf{r}_i, L\mathbf{p}_{i-1}\rangle^2}{|L\mathbf{p}_{i-1}|^2}$$

using the definition (18) of b^i_j.

Proof of Proposition 1 From (3,4)

$$\langle \mathbf{r}_{i+1}, \mathbf{r}_i \rangle = |\mathbf{r}_i|^2 - \frac{\langle \mathbf{r}_i, L\mathbf{p}_i \rangle^2}{|L\mathbf{p}_i|^2} \tag{104}$$

However from (26)

$$|\mathbf{r}_{i+1}|^2 - |\mathbf{r}_i|^2 = -\frac{\langle \mathbf{r}_i, L\mathbf{p}_i \rangle^2}{|L\mathbf{p}_i|^2} \tag{105}$$

so together

$$\langle \mathbf{r}_{i+1}, \mathbf{r}_i \rangle = |\mathbf{r}_{i+1}|^2$$

and hence

$$|a_i L\mathbf{p}_i|^2 \equiv (\mathbf{r}_{i+1} - \mathbf{r}_i)^2 = |\mathbf{r}_{i+1}|^2 + |\mathbf{r}_i|^2 - 2\langle \mathbf{r}_{i+1}, \mathbf{r}_i \rangle = |\mathbf{r}_i|^2 - |\mathbf{r}_{i+1}|^2 \tag{106}$$

Proof of 5.6 By (105) $|\mathbf{r}_{mk+1}| = |\mathbf{r}_{mk}|$ implies

$$\langle \mathbf{r}_{mk}, L\mathbf{p}_{mk} \rangle = 0 \tag{107}$$

so in this case by (21)

$$\langle \mathbf{r}_{mk}, LP_{mk}^{-1}\mathbf{r}_{mk} \rangle = 0 \tag{108}$$

Suppose that for some $m' \geq m$

$$\mathbf{r}_{m'k} = \mathbf{r}_{mk} \text{ and } \mathbf{p}_{m'k} = \lambda_{m'k} P_{mk}^{-1} \mathbf{r}_{mk} + \mu_{m'k} \mathbf{p}_{mk} \tag{109}$$

for some scalars $\lambda_{m'k}$, $\mu_{m'k}$ (as is vacuously true for m'=m). Then from (4,107,108,109)

$$a_{m'k} = 0$$

and hence by (3)

$$\mathbf{r}_{m'k+1} = \mathbf{r}_{m'k} = \mathbf{r}_{mk}$$

and by (17)

$$\mathbf{p}_{m'k+1} = P_{mk}^{-1} \mathbf{r}_{mk} + b_{m'k}^{m'k} (\lambda_{m'k} P_{mk}^{-1} \mathbf{r}_{mk} + \mu_{m'k} \mathbf{p}_{mk}) = \lambda_{m'k+1} P_{mk}^{-1} \mathbf{r}_{mk} + \mu_{m'k+1} \mathbf{p}_{mk}$$

some $\lambda_{m'k+1}$, $\mu_{m'k+1}$. Similarly for each i=m'k+1,...,(m'+1)k+1

$$\mathbf{r}_i = \mathbf{r}_{mk} \; and \; \mathbf{p}_i = \lambda_i P_{mk}^{-1} \mathbf{r}_{mk} + \mu_i \mathbf{p}_{mk}$$

some λ_i, μ_i and hence (109) holds for all m' $\geq$m and the claim follows by induction.

Proof of 5.7

(a) By (106) $|a_i L\mathbf{p}_i| \to 0$ as i $\to \infty$, and (27) is then an immediate consequence of (4,3). (28) then follows by (21). (29,30) follow since $|L\mathbf{p}_i|$ is (by virtue of (25) and §5.3) bounded above.

(b) From (18)

$$-b_{mk+j}^{mk+i} = \frac{\langle LP^{-1}(\mathbf{r}_{mk+i+1} - \mathbf{r}_{mk+j}) + LP^{-1}\mathbf{r}_{mk+j}), L\mathbf{p}_{mk+j} \rangle}{|L\mathbf{p}_{mk+j}|^2} =$$
$$\frac{\langle LP^{-1}(\mathbf{r}_{mk+i+1} - \mathbf{r}_{mk+j}), L\mathbf{p}_{mk+j} \rangle}{|L\mathbf{p}_{mk+j}|^2} + \frac{\langle LP^{-1}\mathbf{r}_{mk+j}, L\mathbf{p}_{mk+j} \rangle}{|L\mathbf{p}_{mk+j}|^2}$$

As m$\to \infty$ $\mathbf{r}_{mk+i+1} - \mathbf{r}_{mk+j} \to \mathbf{0}$ for all $0 \leq$ j $\leq$ i $\leq$ k-1 by Proposition 1, so the penultimate term tends to $\mathbf{0}$ as we are supposing $|L\mathbf{p}_j|$ is uniformly bounded below. The last term is 1 by (24).

C Calculating the Eigenvalues of L and LP^{-1} in the PF and Adjoint Models

Arnoldi-type Methods were used to calculate the extremal eigenvalues shown in Figures 6, 7, 3 etc. If the eigenvalues of L lie in some region of the complex plane an Arnoldi Method will generate a sequence of matrices H_m (where m is the number of iterations used) whose eigenvalues approximate the eigenvalues lying on the periphery of the region (so for example if L is symmetric with positive and negative real eigenvalues the furthest to the right and furthest to the left). For a detailed description and basic properties see Section VI.2 of [6] and for convergence properties see Section VI.7 of [6].

C.1 Extremal Eigenvalues of L

For largest positive or negative eigenvalues of L and LP^{-1} the method we used was essentially that of Algorithm 6.2 of [6].

Choose some $\mathbf{r}$ and set $\mathbf{r}_1$=$\mathbf{r}/|\mathbf{r}|$ then

Do while $j = 1, 2, ..., m$

$$Calculate\, \mathbf{w} \equiv L\mathbf{r}_j \tag{110}$$

Do while $i = 1, 2, ..., j$

$$h_{ij} = \langle \mathbf{w}, \mathbf{r}_i \rangle \tag{111}$$

$$\mathbf{w} \leftarrow \mathbf{w} - h_{ij}\mathbf{r}_i \tag{112}$$

end do while
Set

$$h_{j+1,j} = |\mathbf{w}|$$

$$\mathbf{r}_{j+1} = \mathbf{w}/|\mathbf{w}|$$

end do while

Algorithm 1 *Basic Arnoldi Algorithm used for extremal eigenvalues of L*

This algorithm yields an orthonormal set ($\mathbf{r}_1$,...,$\mathbf{r}_m$) spanning the Krylov subspace

$$span\{\mathbf{r}_1, L\mathbf{r}_1, ..., L^{m-1}\mathbf{r}_1\}$$

and an m × m matrix

$$H_m = \begin{pmatrix} h_{11} & h_{12} & h_{13} & .. & h_{1m-1} & h_{1m} \\ h_{21} & h_{22} & h_{23} & .. & h_{2m-1} & h_{2m} \\ 0 & h_{32} & h_{33} & .. & h_{3m-1} & h_{3m} \\ 0 & 0 & h_{43} & .. & h_{4m-1} & h_{4m} \\ 0 & 0 & 0 & .. & .. & .. \\ 0 & 0 & 0 & .. & h_{mm-1} & h_{mm} \end{pmatrix} \tag{113}$$

Suppose the eigenvalues and normalised eigenvectors of H_m are $\lambda_1^{(m)}$,...,$\lambda_m^{(m)}$ and $\mathbf{y}_1^{(m)}$,...,$\mathbf{y}_m^{(m)}$, set

$$V_m = (\mathbf{r}_1, ..., \mathbf{r}_m)$$

and

$$\mathbf{u}_i^{(m)} = V_m \mathbf{y}_i^{(m)}$$

Then Proposition 6.8 of [6] shows that $\lambda_i^{(m)}$ and $\mathbf{u}_i^{(m)}$ approximate the eigenvalues and eigenvectors of L in the following sense:

$$(L - \lambda_i^{(m)} I)\mathbf{u}_i^{(m)} = h_{m+1,m}(\mathbf{y}_i^{(m)})_m \mathbf{r}_{m+1} \tag{114}$$

where

$$(\mathbf{y}_i^{(m)})_m$$

is the mth (ie last) component of $\mathbf{y}_i^{(m)}$, and hence

$$|(L - \lambda_i^{(m)} I)\mathbf{u}_i^{(m)}| = |h_{m+1,m}(\mathbf{y}_i^{(m)})_m| \tag{115}$$

Since the $\mathbf{r}_1$,...,$\mathbf{r}_m$ are orthonormal and the $\mathbf{y}_i^{(m)}$ are normalised

$$|\mathbf{u}_i^{(m)}|^2 = \mathbf{u}_i^{(m)T}\overline{\mathbf{u}_i^{(m)}} = \mathbf{y}_i^{(m)T} V_m^T V_m \overline{\mathbf{y}_i^{(m)}} = \mathbf{y}_i^{(m)T}\overline{\mathbf{y}_i^{(m)}} = 1$$

Hence the relative error

$$\frac{|(L-\lambda_i^{(m)}I)\mathbf{u}_i^{(m)}|}{|\lambda_i^{(m)}\mathbf{u}_i^{(m)}|} = \frac{|h_{m+1,m}(\mathbf{y}_i^{(m)})_m|}{|\lambda_i^{(m)}|} \tag{116}$$

As a double check we usually also calculated the left hand side of (116) directly when $\lambda_i^{(m)}$ was real. Whenever the relative error was worse than 10% the approximate eigenvalues were discarded.

For eigenvalues of L we used a value of m of 100 and of the 100 values $\lambda_i^{(m)}$ typically 60-90 had a relative error of better than 10%, with as expected (see Section VI.7 of [6]) the most extremal ones having the smallest error (both relatively and absolutely). At the longer timesteps (ie approaching 3600s) as many as 25-30 had completely converged (ie had a relative error of 0.0) to machine precision.

C.2 Extremal Eigenvalues of LP^{-1}

For eigenvalues of LP^{-1} (where P^{-1}=a preconditioner such as XZ ADI etc) (110) in Algorithm 1 was replaced by

$$Calculate\ \mathbf{w} \equiv LP^{-1}\mathbf{r}_j$$

m was set to 60 of which typically just over half had a relative error of better than 10%.

C.3 Smallest Eigenvalues of L

To extract the eigenvalues of L of smallest magnitude we calculated the extremal eigenvalues of L^{-1}, whose eigenvalues are of course the inverse of the eigenvalues of L, by replacing (110) by

$$Solve\ L\mathbf{w} = \mathbf{r}_j$$

which is of course a far more expensive operation than (110) since it involves running GCR(k). Let GCR(k)'s approximate solution $\mathbf{x}$ to

$$L\mathbf{x} = \mathbf{f}$$

be denoted $\tilde{L}\mathbf{f}$. Let $\mu_1^{(m)}$,...,$\mu_m^{(m)}$ be the eigenvalues and $\mathbf{z}_1^{(m)}$,...,$\mathbf{z}_m^{(m)}$ the normalised eigenvectors of H_m obtained from Algorithm 1 with L replaced by $\tilde{L}$, and set

$$W_m = (\mathbf{r}_1, ..., \mathbf{r}_m)$$

and

$$\mathbf{v}_i^{(m)} = W_m\mathbf{z}_i^{(m)}$$

Our estimates of the smallest eigenvalues of L are

$$\lambda_i^{(m)} = \frac{1}{\mu_i^{(m)}}$$

so since as above $|\mathbf{v}_i^{(m)}|$=1 the relative error

$$\frac{|(L-\lambda_i^{(m)}I)\mathbf{v}_i^{(m)}|}{|\lambda_i^{(m)}\mathbf{v}_i^{(m)}|} = \frac{|(L-\lambda_i^{(m)}I)\mathbf{v}_i^{(m)}|}{|\lambda_i^{(m)}|}$$

$$= |\mathbf{v}_i^{(m)} - \mu_i^{(m)}L\mathbf{v}_i^{(m)}| = |(I - L\tilde{L} + L\tilde{L} - \mu_i^{(m)}L)\mathbf{v}_i^{(m)}| \le |(I-L\tilde{L})\mathbf{v}_i^{(m)}| + |L(\tilde{L}-\mu_i^{(m)}I)\mathbf{v}_i^{(m)}| \tag{117}$$

From (114) we know that

$$(\tilde{L}-\mu_i^{(m)}I)\mathbf{v}_i^{(m)} = h_{m+1,m}(\mathbf{z}_i^{(m)})_m\mathbf{r}_{m+1} \tag{118}$$

and hence the last term in the right in (117)

$$|L(\tilde{L}-\mu_i^{(m)}I)\mathbf{v}_i^{(m)}| = |h_{m+1,m}(\mathbf{z}_i^{(m)})_m L\mathbf{r}_{m+1}| \tag{119}$$

As for the penultimate term of (117),

$$(I - L\tilde{L})\mathbf{v}_i^{(m)} = (I - L\tilde{L})(\mathbf{r}_1, ..., \mathbf{r}_m)\mathbf{z}_i^{(m)} = (\mathbf{r}_1 - L\tilde{L}\mathbf{r}_1, ..., \mathbf{r}_m - L\tilde{L}\mathbf{r}_m)\mathbf{z}_i^{(m)}$$

so

$$|(I - L\tilde{L})\mathbf{v}_i^{(m)}| \leq \sum_{j=1}^{m} |(\mathbf{r}_j - L\tilde{L}\mathbf{r}_j)|(\mathbf{z}_i^{(m)})_j| \tag{120}$$

Combining (117,119,120) we have

$$\frac{|(L - \lambda_i^{(m)} I)\mathbf{v}_i^{(m)}|}{|\lambda_i^{(m)}\mathbf{v}_i^{(m)}|} \leq |h_{m+1,m}(\mathbf{z}_i^{(m)})_m L\mathbf{r}_{m+1}| + \sum_{j=1}^{m} |(\mathbf{r}_j - L\tilde{L}\mathbf{r}_j)|(\mathbf{z}_i^{(m)})_j| \tag{121}$$

GCR provides us with the residual error norm $|(\mathbf{r}_j - L\tilde{L}\mathbf{r}_j)|$ for each j, hence (121) enables us to calculate the relative accuracy of the eigenvalues obtained. As a double check we usually also calculated the left hand side of (121) directly when $\mu_i^{(m)}$ and hence $\lambda_i^{(m)}$ was real.

Typically we used m=30, then restarted using as our initial guess (ie $\mathbf{r}_1$) in the second loop (also with m=30) the eigenvector corresponding to the smallest eigenvalue found in the first loop.

C.4 Smallest Eigenvalues of LP^{-1}

To extract the smallest eigenvalues of LP^{-1} (where P^{-1} is a preconditioner such as XZ ADI etc) we need to replace (110) by

$$Solve\ LP^{-1}\mathbf{w} = \mathbf{r}_j$$

Solving equations of this form can be achieved by modifying the preconditioned GCR(k) algorithm of Appendix D, to obtain the following algorithm:

For any initial guess $\mathbf{x}_0$ set

$$\mathbf{r}_0 = L(\mathbf{x}_0) - \mathbf{f}$$

$$\mathbf{p}_0 = \mathbf{r}_0$$

calculate $LP^{-1}(\mathbf{p}_0)$ and defining convergence as the error $\|\mathbf{r}_n\| < \epsilon$ for some $\epsilon \geq 0$
then

Do while $i = 1, 2, \ldots,$ *until convergence or maximum user requested iterations reached*

Do while $v = 0, \ldots, k-1$ *and convergence not obtained*

$$\beta_i = -\frac{\langle \mathbf{r}_i, LP^{-1}(\mathbf{p}_i)\rangle}{\langle LP^{-1}(\mathbf{p}_i), LP^{-1}(\mathbf{p}_i)\rangle}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \beta_i \mathbf{p}_i$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \beta_i L(\mathbf{p}_i)$$

if $\|\mathbf{r}_{i+1}\| < \epsilon$ *then exit*

$$\mathbf{q} = \mathbf{r}_{i+1}$$

Calculate $LP^{-1}(\mathbf{q})$

$$\forall l = 0, v \;\; \alpha_i^j = -\frac{\langle LP^{-1}(\mathbf{q}), LP^{-1}(\mathbf{p}_j)\rangle}{\langle LP^{-1}(\mathbf{p}_j), LP^{-1}(\mathbf{p}_j)\rangle}$$

$$\mathbf{p}_{i+1} = \mathbf{q} + \sum_{l=0}^{i} \alpha_i^j \mathbf{p}_j$$

$$LP^{-1}(\mathbf{p}_{i+1}) = LP^{-1}(\mathbf{q}) + \sum_{l=0}^{i} \alpha_i^j LP^{-1}(\mathbf{p}_j)$$

end do while

set $\mathbf{x}_0 = \mathbf{x}_k$, $\mathbf{r}_0 = \mathbf{r}_k$, $\mathbf{p}_0 = \mathbf{p}_k$, *and* $LP^{-1}(\mathbf{p}_0) = LP^{-1}(\mathbf{p}_k)$
end do while

Algorithm 2 *Algorithm to solve* $LP^{-1}\mathbf{x}=\mathbf{f}$

When using Arnoldi to find the smallest eigenvalues of LP^{-1} we normally used a single loop and m=40, with around a quarter of the eigenvalues obtained being better than 10% accurate.

D The Preconditioned MO GCR(k) Algorithm

For any initial guess $\mathbf{x}_0$ set

$$\mathbf{r}_0 = L(\mathbf{x}_0) - \mathbf{f} \tag{122}$$

$$\mathbf{p}_0 = P^{-1}(\mathbf{r}_0) \tag{123}$$

calculate $L(\mathbf{p}_0)$ and defining convergence as the error $\|\mathbf{r}_n\| < \epsilon$ for some $\epsilon \geq 0$
then
Do while $m = 1, 2, \ldots,$ *until convergence or maximum user requested iterations reached*
Do while $i = 0, \ldots, k-1$ *and convergence not obtained*

$$\beta_i = -\frac{\langle \mathbf{r}_i, L(\mathbf{p}_i)\rangle}{\langle L(\mathbf{p}_i), L(\mathbf{p}_i)\rangle} \tag{124}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \beta_i \mathbf{p}_i \tag{125}$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \beta_i L(\mathbf{p}_i) \tag{126}$$

if $\|\mathbf{r}_{i+1}\| < \epsilon$ *then exit*

$$\mathbf{q} = P^{-1}(\mathbf{r}_{i+1}) \tag{127}$$

Calculate $L(\mathbf{q})$

$$\forall j = 0, i \;\; \alpha_i^j = -\frac{\langle L(\mathbf{q}), L(\mathbf{p}_j)\rangle}{\langle L(\mathbf{p}_j), L(\mathbf{p}_j)\rangle} \tag{128}$$

$$\mathbf{p}_{i+1} = \mathbf{q} + \sum_{j=0}^{i} \alpha_i^j \mathbf{p}_j \tag{129}$$

$$L(\mathbf{p}_{i+1}) = L(\mathbf{q}) + \sum_{j=0}^{i} \alpha_i^j L(\mathbf{p}_j) \tag{130}$$

end do while
set $\mathbf{x}_0 = \mathbf{x}_k$, $\mathbf{r}_0 = \mathbf{r}_k$, $\mathbf{p}_0 = \mathbf{p}_k$, *and* $L(\mathbf{p}_0) = L(\mathbf{p}_k)$
end do while

E Note on the code and settings used for the test runs

The code used was essentially Var Stable 1.5, including the uncorrected implicit weights for the w equation in the PF Model and Adjoint, but with the the bug affecting the block vertical and ADI preconditioner coefficients at levels 2 and model_levels-1 fixed. Unless otherwise specified the defaults for the runs were as follows:

December 1998 linearisation state created by ND Version 2.6a without physics

No lower limit imposed on G (ie no G_term_tol) except for Figure 10

Climate resolution, ie 96 points E-W × 73 rows N-S × 38 model levels

Single timestep of 1800s in Figures 1 and 2

GCR restart values (ie 'k') of 2 in both PF Model and Adjoint, except in Figures 1 and 2

Block vertical preconditioning on the 1st iteration and XZ ADI preconditioning with two pseudo-timesteps thereafter (ie 'option 4') in Figures 1, 2, 16 and 18

The ADI preconditioner used the 'fast' option in x direction (ie L_gcr_fast_x=.true.), starting from zero first guess (ie GCR_zero_init_guess=.true.) without using full operator on right hand side (ie GCR_ADI_ADD_FULL_SOLN=.false.)

ADI pseudo timestep of 0.006s in PF Model, ADI pseudo timestep of

$$-150000s \times \frac{\text{real timestep}}{1800}$$

in Adjoint (ie L_gcr_calc_ADI_ts_Adj=.true.)

The GCR(k) iteration was stopped when infinity norm descended to 10^{-12} in PF Model and 10^{-24} in Adjoint in Figures 15, 16, 17 and 18

References

[1] Owe Axelsson 'Iterative Solution Methods' 1994

[2] Stanley C Eisenstat, Howard C Elman and Martin H Schultz, 1983 'Variational Iterative Methods For Nonsymmetric Systems of Linear Equations' SIAM J Numer Analysis, 20, 345-357

[3] Eltgroth, P G 1994 'Two Portable Parallel Tridiagonal Solvers' in Proc 4th CHAMMP Wksp Argonne 1994

[4] Forsythe, G E and Wasow, W R, Finite-difference methods for partial differential equations, John Wiley 1960

[5] Lancaster and Tismenetsky, The Theory of Matrices, 1985

[6] Saad, Y, 'Numerical Methods for Large Eigenvalue Problems', 1991

[7] Saad, Y (1996), "Iterative Methods for Sparse Linear Systems"

[8] Saad, Y and Schultz, M H (1985), "Conjugate Gradient-Like Algorithms for Solving Nonsymmetric Linear Systems", Math Comp 44, pp 417-424

[9] Skamarock, W C et al, 1997 'Preconditioned Conjugate-Residual Solvers for Helmholtz Equations in Nonhydrostatic Models' MWR Vol 125, p 587

[10] Smolarkiewicz, P K and Margolin 1994 Variational elliptic solver for atmospheric applications Los Alamos Nat Lab Report

[11] Smolarkiewicz, P K et al 'Forward-in-time differencing for fluids: nonhydrostatic modeling of fluid motions on a sphere', 1998, in Proceedings of Seminars 'Recent Developments in Numerical Methods for Atmospheric Modelling' ECMWF, 7-11 Sept 1998

[12] Staniforth, A et al, New Dynamics Documentation for UM5.3, Met Office 2002

[13] Var Science Documentation Paper 52 Version 1.15, 2002

[14] Vinsome, P K W 1976 'Orthomin, an iterative method for solving sparse sets of simultaneous linear equations', in Proc. Fourth Symposium on Reservoir Simulation, Society of Petroleum Engineers of AIME, pp 149-159